Collegio Carlo Alberto

Identity, Dignity and Taboos: Beliefs as Assets

Roland Bénabou
Jean Tirole

Working Paper No. 50
July 2007
www.carloalberto.org

Identity, Dignity and Taboos: Beliefs as Assets[1]

Roland Bénabou[2]
Princeton University

Jean Tirole
Université de Toulouse and MIT

June 2007[3]
(Ealier Draft: December 2006)

[1]We are grateful for helpful comments and suggestions on this project to Andrew Caplin, Robert Oxoby, Philipp Sadowski and Glen Weyl, as well as to participants at several conferences and seminars. Bénabou gratefully acknowledges support from the National Science Foundation and the Canadian Institute for Advanced Research.

[2]CEPR, NBER, and IZA.

Abstract

We analyze social and economic phenomena involving beliefs which people value and invest in, for affective or functional reasons. Individuals are at times uncertain about their own deep values and infer them from their past choices, which then come to define who they are. Identity investments increase when information is scarce or when a greater endowment of some asset (wealth, career, family, culture) raises the stakes on viewing it as valuable (escalating commitments). Taboos against transactions or the mere contemplation of trade-offs arise to protect fragile beliefs about the priceless value of certain assets (life, freedom, love, faith) or things one would never do. Whether such behaviors are welfare-enhancing or reducing depends on whether beliefs are sought for a functional value (sense of direction, self-discipline) or for mental consumption motives (self-esteem, anticipatory feelings). Escalating commitments can thus lead to a hedonic treadmill, and competing identities cause dysfunctional failures to invest in high-return activities (education, adapting to globalization, assimilation), or even the destruction of productive assets. In social interactions, norm violations elicit a forceful response (exclusion, harassment) when they threaten a strongly held identity, but further erode morale when it was initially weak. Concerns for pride, dignity or wishful thinking lead to the inefficient breakdown of Coasian bargaining even under symmetric information, as partners seek to self-enhance and shift blame by turning down insultingly low offers.

Keywords: identity, self-serving beliefs, self-image, memory, wishful thinking, anticipatory utility, self control, hedonic treadmill, inefficient bargaining, taboos, religion.

JEL numbers: D81, D91, Z13.

> Man naturally desires... not only praise, but praiseworthiness; or to be that thing which, though it should be praised by nobody, is, however, the natural and proper object of praise. He dreads, not only blame, but blame-worthiness; or to be that thing which, though it should be blamed by nobody, is, however, the natural and proper object of blame.
> (Adam Smith, *The Theory of Moral Sentiments*)

> A pay cut also represents a lack of recognition. This is true of anybody. People never understand and don't want to understand. They don't want to believe that the company is in that much trouble. They live in their own world and make very subjective judgments.
> (Small business owner, in Truman Bewley, *Why Wages Don't Fall During a Recession*)

Introduction

Many social and economic phenomena involve beliefs which people value and invest significant resources in pursuing, maintaining and defending. The desire to think of oneself as a moral person is a powerful motivator to help others and refrain from cheating, free-riding or consuming certain products. Upholding faith in an afterlife or divine justice requires conforming to rituals, rehearsing sacred texts and abstaining from proscribed behaviors, or even thoughts. Maintaining one's dignity demands that one turn down "insulting" offers that could profitably be accepted, refuse "charity", and fight to defend one's honor or that of the clan. A number of recent experiments similarly document how subjects incur costs and forego decision-relevant information in order to preserve favorable self-concepts relative to their health, fairness or honesty.

This paper aims to analyze, within a unified framework, this broad range of behaviors. The proposed theory is cognitive, in that it models identity and related concepts as beliefs about one's deep "values" and emphasizes the *self-inference* process through which they operate. At the same time, the *needs served* by particular beliefs are linked to more basic aspects of preferences. This "demand side" may reflect a quest for affective benefits, functional ones, or both. The first case arises when self-image has hedonic value or when the future prospects implied by one's economic and social assets give rise to anticipatory utility. The second obtains when a strong sense of self provides clear priorities and directions that help mobilize energy and resist temptations. On the "supply side" of motivated beliefs, the pivotal role is played by imperfect memory (or awareness), which naturally gives rise to identity investments as self-signals: because people have better, more objective access to the record of their conduct than to the exact mix of motivations driving them, they are led to judge themselves (and their situation) by what they do.[1] When contemplating choices, they then take into account what kind of a person each alternative would "make them" and the desirability of those self-views –a form of rational cognitive dissonance reduction.

[1] See, e.g., Festinger and Carlsmith (1959) on cognitive dissonance and especially Bem (1972) on self-perception. On the self-manipulation of "diagnostic" actions see Quattrone and Tversky (1984), and on the strategic management of self-image, see Dana et al. (2003) and Mazar et al (2006).

The first half of the paper develops the basic framework and some general propositions, which it then relates to the experimental evidence. Three main positive results emerge. First, identity investments are higher in situations where objective information is scarce, and conversely they are easily affected by minor manipulations of salience and attention. Second, the model explains *escalating commitments* (Staw (1956)), in which someone who has built up enough of some economic or social asset (wealth, career, family, culture, etc.) continues to invest in it even when the marginal return no longer justifies it. Intuitively, a higher stock raises the stakes on viewing the asset as beneficial to one's long-run welfare, and the way to "demonstrate" such values or prospects is to keep investing. This self-justification leads to excessive specialization (e.g., work versus family) and persistence in unproductive tasks.

Third, identity investment is hill-shaped with respect to the strength of prior beliefs, being highest when people are most uncertain of their long-run values: adolescents, immigrants, new converts, traditional societies faced with globalization. This non-monotonicity also predicts a distinctive pattern of responses to identity threats which can help reconcile a number of divergent experimental findings: whereas challenges to a weakly held identity (low prior) elicit *conformity* effects, effective challenges to a strongly held one (high prior) elicit forceful *counterreactions* aimed at restoring the threatened beliefs. The latter is common with religious and sexual identity (e.g., Mass et al. (2003)); it also corresponds, for moral identity, to the "transgression-compliance" effect, whereby people who are led to believe that they have harmed someone show an increased willingness to perform good deeds (Carlsmith and Gross (1969)). Confirmatory responses to manipulations of an identity that is relatively fragile, on the other hand, correspond to the "foot in the door" effect (accepting a request for a small favor raises the probability of accepting much costlier ones later on; see DeJong (1979), and can also account for the impact of "stereotype threat" on academic performance (Steele and Aronson (1995)).

These positive results are quite general, depending mostly on the "supply" side of the motivated-beliefs mechanism (self-signaling). The welfare consequences of the quest for identity, dignity and similar concerns, on the other hand, depend critically on whether the "demand" side reflects mental-consumption motives (self-esteem, anticipatory utility) or instrumental ones (self-discipline, sense of direction). In the first case, identity investments always reduce expected welfare, being *in fine* a form of wasteful signaling. An individual is then always worse off with malleable beliefs or memory than with non-manipulable ones. Most strikingly, he can even be made worse off by a higher capital stock, as the escalating-commitment mechanism leads to a *treadmill effect* in which higher levels of wealth, social status, or professional achievement induce a self-defeating pursuit of the belief that happiness lies in the accumulation of those same assets. In the second case, by contrast, more malleable beliefs and the resulting ability to shape them through actions can (under specific conditions) raise *ex ante* welfare, by improving the individual's capacity to resist temptations and make consistent choices.

In the second half of the paper, we use the model to examine four main economic applications.

1. Taboos and sacred values. In contrast to economists, most societies and cultures proclaim certain goods to be "priceless" or sacred: life, justice, liberty, love, faith, etc. For such goods, not only are markets often banned as "contrary to human dignity", but even the mere thought of placing a monetary value on them is seen as appalling or sacrilegious. Our model provides an explanation for such "taboo tradeoffs" (Fiske and Tetlock (1997)). We show that upholding certain valued beliefs (or illusions) concerning the "incommensurable" value of certain goods or the things one "would never do" (various forms of "selling out") can require shunning any evaluation, in act or in thought, that might reveal what terms of trade could be obtained. Such *information-aversion* also distinguishes our model of identity from alternatives based on endogenous preferences, in which it cannot arise.

2. Competing identities and oppositional behaviors. When two identities are likely to compete for time or resources in the future, for instance because they entail different lifestyles or locations, investing in one (B) "depreciates" the other (A), as it suggests that the individual may not value it that much. If he has substantial capital vested in A but the long-term value of this asset is more uncertain than that of B (e.g., sentimental or cultural attachments versus easily quantifiable monetary benefits), he may refrain from profitable investments in B and end up worse off. This mechanism can help explain resistance to globalization by traditional societies, or to integration by immigrants and their descendents. It can also take the form of destroying B capital, as with rioting youths burning down their neighborhood schools. Such dysfunctional behaviors become more likely when people turn more pessimistic about their chances of success for investing in B –even though it remains the optimal thing to do– or when the salience of the alternative identity A is amplified by media attention or ideological manipulations.

3. Peer effects and responses to transgressions. Since the preferences and prospects of similar individuals are often correlated (or as long as they are thought to be), "deviant" behavior by peers –violating norms and taboos, fraternizing with outsiders, etc.– conveys *bad news* about the value of the existing capital stock (anticipatory-utility version) or that of motivation-sensitive future investments (imperfect-willpower version). We show that when such transgressions effectively threaten a strong group identity they trigger a forceful response, designed to "repair" the damaged belief. This can involve renewed investment, excluding non-conformers to suppress the undesirable reminders created by their presence, or harassing them. When the initial identity was relatively weak, on the other hand, transgressions will further "sap morale" and depress investment. In both cases, a norm violator's behavior has greater impact, the more similar to the group he is thought to be.

4. Bargaining with malleable beliefs. We show how concerns such as pride, dignity or wishful thinking (anticipatory utility) lead to the inefficient breakdown of Coasian agreements under *symmetric information.* The importance of self-delusion in trials, strikes and other conflicts is

well documented by field observers (e.g., Bewley (1999), Woods et al. (2006)) and experiments (Thompson and Loewenstein (1992), Babcock et al. (1995)). We consider a partnership of two individuals or groups (spouses, capital and labor, majority and minority populations) who must decide whether to continue together or destroy the match. Continuation always yields a positive surplus, but a low output realization means that at least one party has low ability. Moreover, whereas joint output is hard data, individual contributions to it ("who is to blame", "who is getting a raw deal") are soft signals, symmetrically observed when bargaining but imperfectly recalled following a split. Agreeing to inferior or even equal contractual terms in a low-performance team then entails a loss in self image or anticipatory utility. Conversely, by refusing "insultingly low" offers and destroying the match when they do not obtain enough of a concession, each side can try to *shift the blame* onto the other, taking refuge from bleak realities in feelings of self-righteousness or wishful hopes for "a better tomorrow". In equilibrium, the range of sustainable sharing rules is shown to shrink with the importance of self-image or anticipatory concerns. Beyond a point, a bargaining impasse becomes unavoidable, in spite of gains from trade and common knowledge.

The paper relates to two bodies of economic literature. The first one concerns motivated beliefs and self-deception. We unify in the "demand" side of our model mechanisms that are based on a consumption value of beliefs, whether due to anticipatory feelings (Akerlof and Dickens (1982), Loewenstein (1987), Caplin and Leahy (2001), Landier (2000), Brunnermeier and Parker (2005)), a concern for self-image (Köszegi (2004)), or nonlinear moral payoffs (Rabin (1995)), and those that reflect more functional motives (Carrillo and Mariotti (2000), Bénabou and Tirole (2002, 2006a), Battaglini et al. (2005), Dessi (2005)). On the "supply side" of cognition, the role of imperfect memory as the key channel through which belief management operates builds on our earlier work. The combination of anticipatory utility with imperfect recall is also emphasized by Bernheim and Thomadsen (2005), while the idea of self-signaling or self-reputation makes the paper closely related to Bodner and Prelec (2003), Bénabou and Tirole (2004) and Young (2006).[2]

The second body of literature is that on identity (see Davies (2004) and Hill (2006) for surveys). In an influential series of papers, Akerlof and Kranton (2000, 2002, 2005) emphasize how the endogeneity and interdependence of agents' preferences are structured by their choices of a social category, with a wide range of economic implications.[3] In Rabin (1994), Oxoby

[2] On decisions problems with imperfect recall but no demand for motivated beliefs, see Piccione and Rubinstein (1993). The signaling aspect also relates our model to those dealing with social reputation, particularly Bernheim (1994), Austen-Smith and Fryer (2005), Bénabou and Tirole (2006b) and Smith (2006).

[3] Identity is thus represented as an argument in the utility function that depends on the individual's assigned or chosen social category, on the match between (exogenous) "prescriptions" for that category and the individual's given characteristics and behavior, and on his and others' actions. Related models include Shayo (2005) in the context of redistributive politics and Basu (2006) in that of development. On socially interdependent preferences, see also Becker and Murphy (2000). Sen (1985) discusses identity as personal "commitments", which we show

(2003, 2004) and Konow (2000), agents alter their attitudes towards social status or different social norms through costly "dissonance reduction". In Bisin and Verdier (2005), Horst et al. (2005) and Wichardt (2005), preferences evolve across generations through parental investments and evolutionary selection. These three broad approaches and ours are very complementary. Whereas the former share a focus on agents's choices over alternative utility functions (or perceptions represented as preference parameters not directly tied to an information structure), we emphasize the management of beliefs and the cognitive mechanisms through which it occurs. Our model thus endogenizes the identity payoffs and categorical prescriptions in Akerlof-Kranton and related frameworks, as well as the cognitive costs in the second class of models discussed above (we abstract here, on the other hand, from direct preference spillovers). It also leads to distinctive results, such as information-aversion or the fact that being able to manage his own identity can actually make a person worse off. Different cognitive aspects of identity are explored by Fryer and Jackson (2003), who show optimal categorization can lead to ethnic stereotypes, and by Fang and Loury (2005), who model group identity as a shared convention (akin to a language) for the transmission of information.

The paper is organized as follows. Section I presents the model and Section II derives general positive and normative propositions. Four main applications or extensions are then considered: taboos and sacred values in Section III, conflicting identities and resulting dysfunctional behaviors in Section IV, peer effects and responses to transgressions in Section V, and bargaining with dignity concerns in Section VI. Proofs are gathered in the Appendix.

I The Model

"An identity is a definition, an interpretation, of the self... People who have problems with identity are generally struggling with the difficult aspects of defining the self, such as the establishing of long-term goals, major affiliations, and basic values." (Baumeister (1986)).

A Preferences and beliefs

There are three periods, $t = 0, 1, 2$, as illustrated in Figure I. An individual starts at date 0 with an endowment A_0 of some physical or intangible asset which we shall refer to as identity-specific capital. This could be accumulated recognition, wealth, status, good deeds (possibly religion-specific), knowledge of a language or culture, number of friends or children, experiences and memories shared with them, etc. At dates $t = 0, 1$, the individual can "invest" ($a_t = 1$), with return $r_t \geq 0$, or "not invest" ($a_t = 0$). The new capital stock is thus

$$A_{t+1} = A_t + a_t r_t, \tag{1}$$

can be modeled in a way that is consistent with standard (consequentialist) economic rationality.

where, to lighten the notation, we leave implicit the fact when there is depreciation (or shocks), A_t should be adjusted accordingly. The "investment" action plays here a dual role. The first one is standard accumulation: $r_t > 0$ when the stock can be increased (vita, wealth, friends), whereas $r_t = 0$ for an immutable trait (gender, race). The second and more important role is informational: even when $r_t \equiv 0$, an individual's choice will constitute a signal of how much he values the benefits that flow from the asset A.

Indeed, the central ingredient in the model is that the individual is, at times, unsure of his own deep preferences: personal priorities, moral standards, strength of faith, commitment to culture or career, etc. Such uncertainty over "long-term goals, major affiliations, and basic values" (Baumeister) means that the stock A_2 he will eventually consume from may prove to be very valuable to his long-term welfare, or not that meaningful.

• *Date 0*. At the start of period 0, the individual receives a signal (intuitive feeling, conscious self-assessment, external feedback) about his type:

$$v = \begin{cases} v_H & \text{with probability } \rho \\ v_L & \text{with probability } 1-\rho \end{cases}, \tag{2}$$

with $v_H > v_L$ and $\bar{v} \equiv \rho v_H + (1-\rho) v_L$ denoting the prior expectation. Conditional on v, the expected long-run utility to be derived from A_2 is vA_2. Following the signal, the individual makes his investment decision, $a_0 \in \{0, 1\}$, resulting in a flow payoff $U(v, A_0, a_0)$.

Assumption 1 *The instantaneous utility $U(v, A_0, a_0)$ received by the agent at date 0 satisfies $U_{13} \geq 0$ and $U_{23} \geq 0$.*

The condition $U_{13} \geq 0$ allows us to represent the date-0 impact of investment as a type-dependent cost (or benefit if negative), $c_0^i \equiv U(v_i, A_0, 0) - U(v_i, A_0, 1)$ for $i = H, L$, such that

$$c_0^H \leq c_0^L. \tag{3}$$

When $U_{23} = 0$, as will be the case in most of our applications, the costs c_0^i are independent of the initial stock A_0.

• *Date 1.* The individual's perception of his type at $t = 1$ may often differ from what it was at $t = 0$. The usual assumption is that he gains, through experience, better knowledge of his preferences. For a person's past actions to define his sense of identity, however, it must be that he *no longer has direct access* to the deep motives and feelings that gave rise to them –an information *loss.* Otherwise, past behavior conveys no useful information, so there is no sense in which one can make (or claim to make) choices intended to "be true to myself," "maintain my integrity," "keep my self-respect", "stand for my principles," "not betray my values", "be able to look at myself in the mirror," and the like. And indeed, psychologists provide extensive

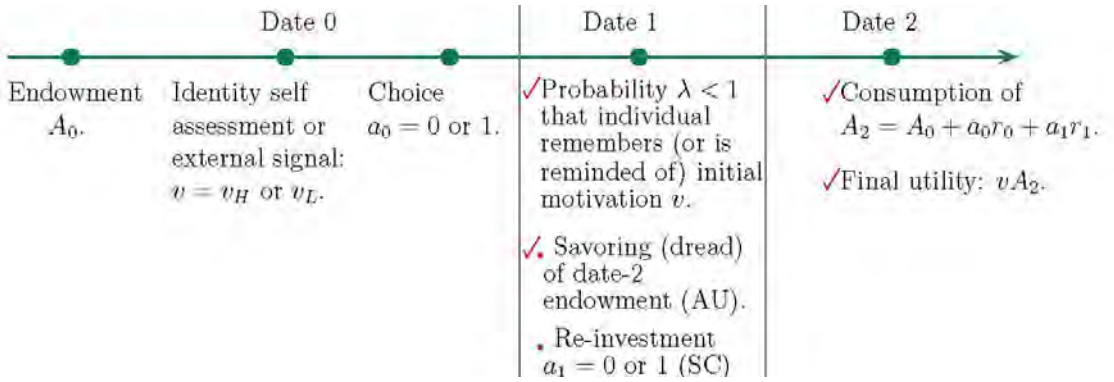

Figure I: Timing of Moves and Actions

evidence that people's recall of their past feelings and true motivations is very imperfect and often self-serving, that they judge themselves by their actions, and that many decisions are shaped by a concern to achieve or maintain desirable self-views.[4]

Assumption 2 *(Self-inference). At date* 1, *the individual is aware (or reminded) of his past motivational state* v *only with probability* λ. *With probability* $1-\lambda$, *he no longer recalls (has access to) it and uses instead his past choice of* a_0 *to infer his type.*

Let us denote by $\hat{\rho}$ the individual's date-1 belief about "what kind of a person" he is and by $\hat{v} \equiv \hat{\rho} v_H + (1-\hat{\rho}) v_L$ the corresponding expected valuation of A_2, either of which defines his (subjective) "sense of identity" at $t = 1$. With probability λ the posterior $\hat{v}$ is thus equal to the actual signal v, and with probability $1-\lambda$ it is equal –with a slight abuse of notation– to the conditional expectation $\hat{v}(a_0) \in [v_L, v_H]$ formed on the basis of previous behavior. More generally, $1-\lambda$ should be thought of as the *malleability of beliefs through actions,* and thus also reflecting the possibility that behaviors may themselves be forgotten or repressed, or be uninformative due to situational factors (e.g., there could be a plausible "excuse").[5]

This process of *self-inference* can be thought of as representing the "supply side" of motivated beliefs in the model. We next turn to the "demand side," which encompasses a number of mechanisms that make certain beliefs more desirable to hold than others. These include pure

[4] On imperfect retrospective and prospective access to feelings and desires, see Kahneman et al. (1997) and Loewenstein and Schkade (1999). On self-perception, see footnote 1. Further discussions and analyses of these phenomena can be found in Bodner and Prelec (2003), Bénabou and Tirole (2004) and Battaglini et al. (2005).

[5] If an action is uninformative with probability ν, the posterior $\hat{v}$ equals v, $\bar{v}$ or $\hat{v}(a_0)$ with respective probabilities λ, $(1-\lambda)\nu$ and $(1-\lambda)(1-\nu)$, so the effect on signaling incentives is similar to that of a decrease in $1-\lambda$. For a model of self-reputation with misremembered actions and excuses, see Bénabou and Tirole (2004). The recall or awareness probability could also be different for good and bad signals, $\lambda_H \geq \lambda_L$, whether exogenously or endogenously (see Bénabou and Tirole (2002). We focus here on the case in which $\lambda_H = \lambda_L$, both for simplicity and to highlight the role of self-inference, which seems most relevant to "identity".

self-image concerns, anticipatory utility and imperfect self-control, all of which can be cast as alternative specifications of the continuation value $V(v, \hat{v}, A_1)$, evaluated at $t = 0$, of entering period 1 with beliefs $\hat{v}$ and capital A_1.

Assumption 3 *The value function* $V = V(v, \hat{v}, A_1)$ *satisfies* $V_2 > 0,\ V_{12} \geq 0$ *and* $V_{13} > 0$.

The first condition is mainly a "good identity" convention.[6] The cross-partial conditions, together with $U_{13} \geq 0$, will generate a sorting condition leading the high-valuation type to always invest at least as much as the low-valuation one, so that actions indeed have informational content.[7] Before analyzing the equilibrium, however, we show how different preferences known to generate a demand for self-serving beliefs map into the value function V. The two main examples are summarized in Figure I.

• *Demand for beliefs 1*: *anticipatory utility (AU) or self-esteem.*

In period 2, the agent obtains from the stock A_2 a utility vA_2. During period 1 he derives from the prospect of that future consumption an anticipatory pleasure or pain $s_1\hat{v}A_2$, where $\hat{v}$ is his date-1 expectation of v and s_1 a "savoring" parameter. An important determinant of s_1 is *salience* –the extent to which the individual thinks (perhaps prompted by an experimenter or advertiser) about the contribution of A_2 to his future welfare.

We focus here on the "pure" anticipatory-utility case, in which there is no further decision to be made at date 1 (Example 3 will incorporate one). Thus $a_1 = 0$, $A_2 = A_1$ and the continuation value of entering period 1 with subjective identity $\hat{v}$ is

$$V(v, \hat{v}, A_1) \equiv (\delta_1 s_1 \hat{v} + \delta_2 v)\, A_1, \tag{4}$$

where δ_1 and δ_2 reflect standard time discounting (back to $t = 0$), with possibly different lengths of periods 1 and 2. Assumption 2 is clearly satisfied, with $V_{13} > 0$, $V_{23} > 0$ and $V_{12} = 0$.

Turning now to date-0 payoffs, let

$$U(v, a_0, A_0) = -ca_0 + \tau v a_0 + s_0 v(A_0 + a_0 r_0), \tag{5}$$

The first term is a time, effort or monetary cost, independent of type. The second one represents consumption benefits derived or "sampled" in the process of investment: socializing with friends, spending time with the family, attending church, fixing up the farm, etc. The third term arises when the agent derives anticipatory utility at $t = 0$, as he does at $t = 1$, from his long-term ($t = 2$) consumption prospects. These last two terms capture intuitive effects that make identity

[6] Furthermore, it will only be used to select the Pareto-dominant equilibrium in the case of multiplicity.

[7] Since a_t and A_t, like v and $\hat{v}$, can always be redefined with the opposite sign, all one really needs is that there exist $(\varepsilon, \eta) \in \{-1, 1\}^2$ such that the functions $U(\varepsilon v, \eta A_0, \eta a_0)$ and $V(\varepsilon v, \varepsilon\widehat{v}, \eta A_1)$ satisfy Assumptions 1 and 3. Note, finally that when $r_0 = 0$, no condition on V_{13} is necessary.

investments less costly, or more pleasant, for the high-valuation type. Thus Assumption 1 is satisfied, with $U_{13} = U_{23} = 0$ and costs $c_0^i = c - (s_0 r_0 + \tau)v_i$ such that $c_0^H < c_0^L$.

Finally, when performing welfare analysis, our criterion will be total intertemporal utility

$$W \equiv E[U + V], \tag{6}$$

where the expectation is taken with respect to the prior distribution $(\rho, 1-\rho)$ of types $v \in \{v_H, v_L\}$ and the (endogenous) distribution $(\lambda, 1-\lambda)$ of posterior beliefs $\hat{v} \in \{v, \hat{v}(a_0)\}$.

This simple benchmark easily accommodates a number of extensions.

a) Self-image or utility from memories. Pure "mental consumptions" (Schelling (1985)) are a special case in which there is no final date at which the true signal v is directly (re-) experienced. Thus $\delta_2 = 0$, and hence $V(v, \hat{v}, A_1) = \delta_1 \hat{v} A_1$, corresponds to an individual who cherishes memories of how honest, productive, or generous he has been. If the stock corresponds to a fixed trait ($A_1 = A_0$), moreover, this specification (or any nonlinear variant) captures a pure demand for self-esteem –with respect to intelligence, attractiveness, and the like.

b) Disappointment aversion. Whereas savoring provides a motive to be optimistic about the future, the fear of being disappointed when consumption eventually occurs generates an opposing incentive to maintain low expectations. Let $S(v, \hat{v}, A_1) = \delta_2 \varphi((v - \hat{v})A_1)$ represent the corresponding period-2 payoff, where φ is increasing, concave and such that $-x\varphi''(x)/\varphi'(x) < 1$ for all x. Concavity, which means that negative surprises weigh more than positive ones (see Gul (1991)), implies $S_{12} > 0$, while the elasticity condition ensures that $S_{13} > 0$ nonetheless. Thus, adding this term into the continuation value V only reinforces the sorting condition in Assumption 2, while generating a demand for "defensive pessimism".[8]

c) Wishful thinking impairs later decisions. The savoring motive will lead agents to distort their initial ($t = 0$) actions in the pursuit of more pleasant beliefs. Once beliefs have been altered, moreover, any subsequent decision-making will also be impacted; such will the case in Example 3 below, with the main difference being that the value function becomes nonlinear.[9]

• *Demand for beliefs 2: imperfect self-control (SC)*

Whereas individuals with anticipatory or self-esteem preferences want to hold certain beliefs for purely *affective* reasons, having a strong, stable sense of identity (or of divine justice) can also be valuable for making consistent choices and persevering in long-term projects. This *functional* motive, equally stressed by psychologists, leads to our second main benchmark.

[8] For V_2 to remain positive, this effect must not be too strong relative to that generated by s_1. Alternatively, it could be so strong as to make V_2 negative everywhere; see footnote 7. What is needed is that $\delta_1 s_1 v + \delta_2 \varphi((\hat{v} - v)A_1)$ be monotonic in v over all feasible values of $v, \hat{v}$ and A_1.

[9] The hedonic value of beliefs in period 1 could also be nonlinear in probabilities. Our *positive* results (Propositions 1 and 2) apply as well to such date-1 payoffs $\pi(\hat{v}, A_1)$, as long as $\pi_1 > 0$ and $\pi_{12} > 0$. As Propositions 3 and 4 make clear, however, *normative* conclusions depend importantly on linearity or the specific form of nonlinearity.

As before, the stock A_2 generates consumption benefits vA_2 at date 2, but now accumulation can take place both at $t = 0$ and at $t = 1$. The latter involves a cost c_1 (which, for simplicity, we take to be type-independent),[10] with

$$\delta_2 v_L r_1 > \delta_1 c_1, \tag{7}$$

where δ_1 and δ_2 have the same interpretation as above. Ex-ante, it is therefore always efficient to invest at $t = 1$, even for someone with relatively low valuation for the identity-related good. Come date 1, however, *weakness of will* can make the immediate disutility of effort much more salient than the distant benefit, giving rise to a self-control problem. Let the individual's "Self 1" thus perceive the current cost as equal to c/β_1, where the willpower (time-consistency) parameter β_1 is drawn at $t = 1$ from a continuous distribution F on $[0, 1]$.[11] Given a posterior belief $\hat{v}$, the individual invests at $t = 1$ only if

$$\beta_1 \delta_2 \hat{v} r_1 \geq \delta_1 c_1, \tag{8}$$

which defines a cutoff level of β_1 that decreases with $\hat{v}$. The continuation value is thus

$$V(v, \hat{v}, A_1) \equiv \delta_2 v A_1 + (\delta_2 v r_1 - \delta_1 c_1) \left[1 - F\left(\frac{\delta_1 c_1}{\delta_2 \hat{v} r_1}\right)\right], \tag{9}$$

which again satisfies all the conditions of Assumption 2. In period 0, finally, let the payoff U, as perceived contemporaneously (i.e., by "Self 0"), be defined as in (5) but with $s_0 = 0$, resulting in net investment costs $c_0^H \leq c_0^L$ that again satisfy Assumption 1.

With regard to welfare analysis, it may no longer be appropriate to just add up $E[U]$ and $E[V]$, as the individual may suffer from present–biased preferences at date 0, as he does at date 1. Suppose that his perceptions of contemporaneous payoffs are magnified by $1/\beta_0$, where $\beta_0 \leq 1$ measures willpower at $t = 0$ (one could easily make it stochastic, as with β_1). Thus, if c_0 is the perceived investment cost, the "real cost", as viewed by an ex-ante self or parent (at date "-1"), is only $\beta_0 c_0$. Recalling that V is a value function and therefore (unlike U) not subject at date-0 to salience of the present, our welfare criterion will be:

$$W = E[\beta_0 U + V], \tag{10}$$

where, as before, the expectation is taken with respect to the prior distribution of types and the posterior distribution of beliefs.

[10] More generally, it suffices that c_1 either be only imperfectly informative about v, or that the agent need to make the $t = 1$ investment decision before having experienced the full cost.

[11] Alternatively, it could be the date-1 cost c_1 that is unknown at date 0. The role of uncertainty here is only to smooth over $t = 1$ decisions so as to make V differentiable (which we use only to simplify the exposition).

- *Demand for beliefs 3: wishful thinking and procrastination*

When does the desire to indulge in pleasant beliefs and avoid unpleasant ones aggravate the self-control problem, and when does it alleviate it? To answer this question, we simply combine the AU and SC specifications and allow for type-dependent returns in investment. Denote those as $r_t(v)$ and the resulting contributions to final utility vA_2 as $z_t(v) \equiv vr_t(v)$, $t = 1, 2$. For an agent with self-view $\hat{v} \in [v_L, v_H]$, or equivalently $\hat{\rho} \equiv (\hat{v} - v_L)/(v_H - v_L) \in [0, 1]$, the corresponding marginal expected utility is then

$$z_t(\hat{v}) \equiv \hat{\rho} z_t(v_H) + (1 - \hat{\rho}) z_t(v_L). \tag{11}$$

He invests at $t = 1$ if $\beta_1 (\delta_1 s_1 + \delta_2) z_1(\hat{v}) \geq \delta_1 c_1$, leading to

$$V(v, \hat{v}, A_1) \equiv (\delta_1 s_1 \hat{v} + \delta_2 v) A_1 + [\delta_1 s_1 z_1(\hat{v}) + \delta_2 z_1(v) - \delta_1 c_1] \left[1 - F\left(\frac{\delta_1 c_1}{(\delta_1 s_1 + \delta_2) z_1(\hat{v})}\right)\right], \tag{12}$$

which satisfies $V_{12} > 0$ as long as z_1 is strictly monotonic, $V_{13} > 0$, and $V_{23} > 0$ if $s_1 > 0$.[12] More optimistic beliefs $\hat{v}$ enhance savoring of the existing stock A_1, but whether they induce higher investment or "coasting" hinges on whether $z_1(\hat{v})$ rises or falls, generating an intuitive dichotomy between situations in which "identity" and effort are *complements or substitutes.*

a) Wealth accumulation, status-seeking, and other entrepreneurial behaviors (complementarity). When $z_1(v)$ is increasing in v, wishful thinking helps alleviate the motivation problem, if there is one (otherwise, it only results in excessive activism). This case occurs for instance if r_1 is type-independent (financial assets) or if v corresponds to some ability that raises both the probability of winning in a competitive situation and the expected value of the prize. Dreams of riches and glory –and of how enjoyable those will be– thus propel entrepreneurs, explorers, athletes and scientists to sacrifices and persistence in the pursuit of long-term endeavors.

b) Health investments, safe driving and other risk-prevention behaviors (substitutability). In those cases $z_1(v)$ is decreasing in v, which may stand for a favorable genetic endowment that protects from disease and makes taking care of one's health less of a necessity, or good driving skills and reflexes that permit faster speeds.[13] Wishful thinking –understating the likelihood of illness, accident or death– then makes the present more enjoyable but further encourages negligent behaviors –unhealthy lifestyle, addictions, careless driving, failing to save for old age–

[12] If $z_1(v)$ is decreasing, one needs to impose conditions such that V_2 remains positive (over the relevant range). Note also that, in the limiting case in which there is *anticipatory utility* but *no present bias,* the term in $1 - F$ becomes $\mathbf{1}_{\{\delta_1 s_1 + \delta_2 z(\hat{v}) \geq \delta_1 c_1\}}$, which is non-differentiable but retains the key increasing-differences properties. It is then easily seen from (12) that distorted beliefs, $v \neq \hat{v}$, *always* lead to (weakly) *suboptimal* decisions at $t = 1$, namely overinvestment or underinvestment, depending on $z_1(\hat{v}) \gtrless z_1(v)$.

[13] In the health case, for instance, the individual's long-term health is $vA_2 = vA_0 + z_0(v)a_0 + z_1(v)a_1$, where v is his endowment of "good" genes and A_0 a constant that can be normalized to 1. On how denial of death impairs decision-making, see Becker (1973) and Kopczuk and Slemrod (2005).

that are precisely those to which weakness of will already makes one too tempted to succumb.[14]

B Interpreting the model

Before proceeding to solve the model, we point out three important ways in which it is more broadly applicable than what a literal reading of the assumptions might suggest.

Identity as multidimensional. The asset-value pair (A, v) can be any of several independent ones (e.g., morality, health, gender). More interestingly, it can also represent a *tradeoff* between two dimensions A and B, such as career and family, linked by uncertainty over their *relative* value $(v_A - v_B)$ and a resource constraint on total investment. The model is then essentially the same, with everything now interpreted in a "differential" sense, in terms of A relative to B.[15] Section IV, moreover, will explicitly study other types of conflicts between identities.

Identity as socially determined. In Sections V and VI, interactions with others will shape (and respond to) an individual's sense of self. Even in the basic model, however, one should already think of the social environment as a key determinant of initial endowments (wealth, education, race, culture), prior beliefs (optimism, religion, politics) and information flows.

Knowledge and affirmation of values. The assumption that people have imperfect retrospective insights into their own motives and feelings admits several formally equivalent interpretations:

i) An *ego-superego* view, in which v is simultaneously known at the subconscious level and not known at the conscious level (see Bodner and Prelec 2004). This corresponds in the model to a limiting case of "instantaneous forgetting".

ii) A *moral sentiments* view, in which people experience guilt or pride not only when actually observed by others (social signaling), but also from the virtual judgements of imagined "spectators" (Smith (1759)).

iii) The *intergenerational transmission of values.* In this polar limiting case, "forgetting" takes a generation, so the date-0 agent is a parent and the date-1 agent his child. Parents have some experience v about the value of certain asset(s), such as the life satisfaction derived from career versus social bonds, the richness of a culture or the comforts of a religion. Children start

[14] Date-0 payoffs are still specified as in (5), or $c_0^i = c - (s_0 r_0(v^i) + \tau)v_i$ for type $i = H, L$, but with: $\tau \geq 0$ in case (a), capturing again the idea that the consumption value inherent to the activity may be sampled in the process of accumulation; $\tau \leq 0$ in case (b), meaning that investing in (say) health confers more immediate benefits to the low-immunity type. Thus $s_0 z_0(v) + \tau v$ is increasing under (a), satisfying Assumption 1, and decreasing under (b), contributing to a "reverse" sorting condition that will make the v_L type more likely to invest at $t = 0$, as he is at $t = 1$. In the latter case one can just redefine "investment" as $b_t = 1 - a_t$ (see footnote 7).

[15] Taking an extreme case of this resource rivalry, suppose that: i) the agent can invest in either A or B $(a_t = 1 - b_t \in \{0, 1\})$, with respective returns r_{At}, r_{Bt}, salience s_A, s_B, and similar notation for other parameters; ii) his long-term values are subject to a *relative preference shock:* $v_A = \bar{v}_A + v/2$ and $v_B = \bar{v}_B - v/2$, where $v = \varepsilon > 0$ with probability ρ and $v = -\varepsilon$ with probability $1 - \rho$. The model is then essentially isomorphic to the basic one, with all variables redefined as differentials. Thus, the relevant asset is now the row vector $A' \equiv (A - B)$, so that "a higher stock" means a higher A, a lower B or both (with enough parameter symmetry, only the scalar $A - B$ matters, but that need not generally be the case) and similarly for $r' \equiv (r_{At} - r_{Bt})$, $s' \equiv (s_{At} - s_{Bt})$, etc.

less well informed and learn (with probability $1-\lambda$) from what they see their parents do, or force them to do (a_0). Parents strive, altruistically or selfishly, to inculcate in their children "values" (beliefs $\hat{v}$) that will enrich their lifetime experience or lead them to take desirable actions. Although we shall generally not reiterate this interpretation of the model, it is quite important and should be kept in mind throughout the paper.

II Equilibrium and welfare

A Investment behavior

At date 0, each type chooses his action optimally, taking into account the impact that may result for his sense of identity at date 1 and the affective and/or functional payoffs that flow from it. Thus a_0 is a solution to

$$\max_{a_0 \in \{0,1\}} \left\{ U\left(v, A_0, a_0\right) + \lambda V\left(v, v, A_0 + a_0 r_0\right) + (1-\lambda) V\left(v, \hat{v}\left(a_0\right), A_0 + a_0 r_0\right) \right\}, \tag{13}$$

where the posterior beliefs $\hat{v}\left(a_0\right)$ in case of self-inference are derived from Bayes' rule. [16] Denoting by x_H and x_L the probabilities that types v_H and v_L respectively invest at $t=0$, this means that $\hat{v}\left(a_0\right) \equiv \hat{\rho}\left(a_0\right) v_H + \left[1 - \hat{\rho}\left(a_0\right)\right] v_L$, where

$$\hat{\rho}(1) \ = \frac{\rho x_H}{\rho x_H + (1-\rho) x_L} \quad \text{and} \qquad \hat{\rho}(0) = \frac{\rho(1-x_H)}{\rho(1-x_H) + (1-\rho)(1-x_L)} \tag{14}$$

for all (x_H, x_L) not equal to $(0,0)$ and $(1,1)$ respectively. To shorten the notation, let us define the expected value function

$$\mathbf{V}\left(v, \hat{v}, A_1\right) \equiv \lambda V\left(v, v, A_1\right) + (1-\lambda) V\left(v, \hat{v}, A_1\right), \tag{15}$$

which brings together the *demand* (preferences) and *supply* (cognition) sides of the model and inherits from V all the properties in Assumption 3. Investing at $t=0$ is thus an optimal strategy for type $v_i \in \{v_H, v_L\}$ if

$$\mathbf{V}\left(v_i, \hat{v}\left(1\right), A_0 + r_0\right) - \mathbf{V}\left(v_i, \hat{v}\left(0\right), A_0\right) - c_0^i \geq 0. \tag{16}$$

There are three reasons why this net return is greater for the v_H type than the v_L type (a sorting condition), implying that if $x_L > 0$ then $x_H = 1$ (hence $\hat{v}\left(1\right) \geq \hat{v}\left(0\right)$ on the equilibrium path).

[16] The problem we study thus has the structure of a dynamic "psychological game" (Geanakoplos et al. (1989)) between the individual's time 0 and time 1 "selves". By modeling agents as Bayesian, and thus aware that they sometimes make decisions so as to maintain or enhance a valued identity, we are treating them as fairly sophisticated. Relaxing this "metacognition" assumption (e.g., Bénabou and Tirole (2002)) would make the model's positive results only stronger, but lead in certain cases to different welfare implications (see footnote 22).

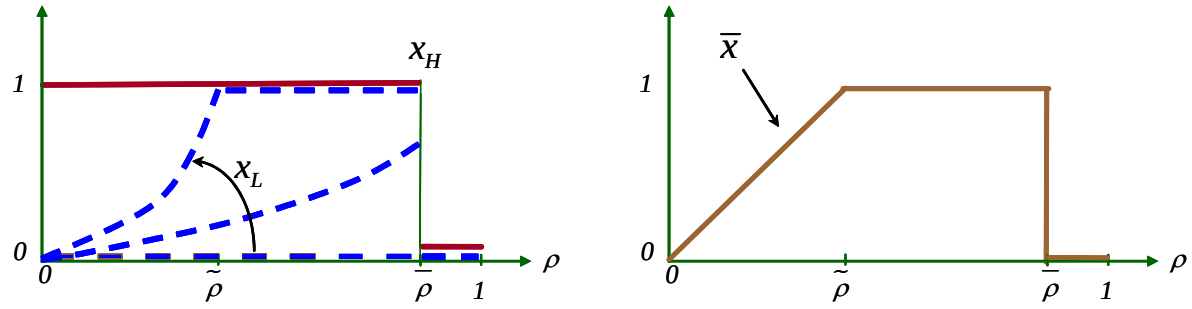

Figure II: Equilibrium as a function of ρ. Left panel: solid line $= x_H(\rho)$, thick dashed line $= x_L(\rho)$, for decreasing values of c_0^L. Right panel: average investment $\overline{x}(\rho)$.

First, the high-valuation type has a lower effective cost, $c_0^H \leq c_0^L$. Second, when $V_{13} > 0$, he attaches greater value to any addition to the capital stock. Finally, when $V_{12} > 0$ he also cares more about having a "strong" identity at date 1, which investing helps achieve if $\hat{v}(1) > \hat{v}(0)$.

From now on, we shall restrict attention to *monotonic equilibria,* defined as those in which: i) the high-value type always invests more: $x_H \geq x_L$, which given (16) again means that $x_H = 1$ whenever $x_L > 0$; ii) a (stronger) form of monotonicity is also imposed on off-the-equilibrium-path beliefs: if $x_H = x_L = 0$, then $\hat{\rho}(1) \equiv 1$; symmetrically, if $x_H = x_L = 1$, then $\hat{\rho}(0) \equiv 0$. This refinement is intuitive and does not affect any qualitative results.[17]

Finally, over a certain range of parameters there may be multiple (three) monotonic equilibria, among which one is Pareto-dominant and will be the one selected.[18]

Proposition 1 *There exists a unique (monotonic, undominated) equilibrium, characterized by thresholds* $\tilde{\rho}$ *and* $\bar{\rho}$ *with* $0 \leq \tilde{\rho} \leq \bar{\rho} \leq 1$ *and investment probabilities* $x_H(\rho)$ *and* $x_L(\rho)$ *such that:*
(1) $x_H(\rho) = 1$ *for* $\rho < \bar{\rho}$ *and* $x_H(\rho) = 0$ *for* $\rho \geq \bar{\rho}$;
(2) $x_L(\rho)$ *is non-decreasing on* $[0, \tilde{\rho}]$, *equal to* 1 *on* $[\tilde{\rho}, \bar{\rho})$ *when* $\tilde{\rho} < \bar{\rho}$ *and equal to* 0 *on* $[\bar{\rho}, 1]$.

The equilibrium is illustrated in Figure II, for the case where $0 < \bar{\rho} < 1$ and for decreasing values of c_0^L, so as to illustrate all the cases of interest:

(i) no investment: when ρ is high enough ($\rho > \bar{\rho}$), the v_H type can afford not to invest, knowing that since the other type also abstains, the posterior will equal the prior, which is already close to 1 and thus could not be increased much anyway.

When initial self-confidence is below the threshold $\bar{\rho}$, on the other hand, the v_H type needs to invest in order to "affirm his values" and separate from the more common v_L type. Turning now to the latter's behavior, one of three cases arises.

[17] It is implied for instance by Cho and Kreps' (1987) Never a Weak Best Response (NWBR) criterion if $V_{12} = 0$ (as is the case for *AU*).

[18] An equilibrium Pareto dominates another one if it yields a weakly higher payoff to both types and a strictly higher payoff to at least one of them.

(ii) separation: when c_0^L is sufficiently high, the low-valuation type never finds it worthwhile to invest ($\tilde{x} = 0$), whereas the high-valuation does, for $\rho < \bar{\rho}$;

(iii) randomization by v_L : for lower values of c_0^L, it becomes desirable for the v_L type to imitate the v_H type, but his ability to do so profitably is limited by the initial prior ($0 < \tilde{x} < 1, \tilde{\rho} = \bar{\rho}$). The lower is ρ, the more truthful (low x_L) his strategy must be in order for investment to signal a high type with sufficient credibility (see (14)).

(iv) full investment: for c_0^L still lower, even a small signaling gain is profitable, so the low-valuation type can afford to completely pool with the other one ($\tilde{x} = 1$), provided ρ is above some threshold $\tilde{\rho}$ (which increases with c_0^L).

Having fully characterized equilibrium behavior, we now derive comparative-statics predictions and relate them to the available experimental evidence. We shall say that the individual invest more in identity if both x_H and x_L (weakly) increase –and hence so does the total probability of investment, $\bar{x} \equiv \rho x_H + (1-\rho)\, x_L$.[19]

Proposition 2 *(1) An individual invests more in identity:*
(i) the more malleable his beliefs (the lower λ),
(ii) the lower the investment cost (the lower c_0^L or c_0^H),
(iii) the more salient the identity in the AU case (higher s_1).
(iv) the higher the capital stock A_0 in the AU case, and more generally when $V_{23} \geq 0$.
(2) Initial beliefs have a nonmonotonic, hill-shaped, effect on overall investment: $\bar{x}$ increases linearly in ρ on $[0, \tilde{\rho})$, equals 1 on $[\tilde{\rho}, \bar{\rho})$, then falls to 0 beyond.

B Implications and experimental evidence

These results can help understand a broad range of empirical phenomena. While some of those admit alternative explanations (such as learning by doing, habit formation, or unstable preferences), a different one would have to be invoked in each case. Our model, by contrast, aims to provide a single account for all of them, as well as for the four economic applications considered in the second part of the paper.

1) Malleability of beliefs. An increase in the probability λ that the individual remains aware, or is reminded of, his true motives and values, reduces investment. Identity-management is thus more likely to occur in settings that are unfamiliar or in which verifiable information is scare (e.g., religion). A more operationalizable source of variation in $1-\lambda$ (discussed following Assumption 2) is the extent to which actions are informative about ones' underlying "character", or could instead be attributed to mistakes, rationalized by situational factors, etc. Dana et al. (2003) document the importance of such inferential *"wriggle room"* for altruistic self-image: when

[19] Given Proposition 1, illustrated in Figure II, the fact that (for all ρ) x_H increases means that $\bar{\rho}$ increases, and the fact that (for all ρ) x_L increases means that either $x_L(\bar{\rho})$ increases or $x_L(\bar{\rho}) = 1$ and $\widetilde{\rho}$ decreases.

subjects in a dictator-like game did not know whether their payoff and that of the recipient were positively or negatively related, but could find out at no cost by clicking on a button, over half of them chose not to know and proceeded to make the self-serving choice, whereas when faced with an explicit tradeoff two-thirds chose a "fair" allocation. Mazar et al. (2006) document a similar effect of attributional ambiguity on self-image investments pertaining to honesty: when subjects whose payment was based on their self-reported, unverifiable performance on a task earned their compensation in the form of tokens that would later on be exchanged for money (at a known rate), the overinflating of claims (assessed relative to a verifiable-performance benchmark) was 50% higher than when they had to lie for cash directly.

2*) Salience of identity.* Messages or cues that "remind" individuals of specific components of their identity will elicit investments along the same dimensions. LeBoeuf and Shafir (2004) thus find that even minor manipulations emphasizing alternative aspects of subjects' self-concept, such as scholar versus socialite, or ethnic Chinese versus American citizen, trigger identity-consistent expressions of consumption preferences. In experiments with monetary stakes, Benjamin et al. (2006) find that similarly priming subjects to their ethnic identity caused Asian-Americans to make considerably more patient choices, Whites to makes choices that were both more patient and less risk averse, and non-immigrant African-Americans to make more risk-averse ones. In Mazar et al. (2006), making personal honesty more salient by having subjects read the Ten Commandments before performing tasks in which they could profitably cheat on their claimed output without risk of detection led to significant decreases in claims inflation.

3) Escalating commitment. The more identity-relevant capital they have, the more identity-affirming investment people will make, thereby raising the stock even further. This result is not due to any increasing returns in the investment technology: in our three benchmark cases, $U_{23} = 0$. The reason is instead that someone with more A_0 has a greater vested interest in viewing this asset as valuable rather than worthless, and further investment is the way to demonstrate such beliefs –as in the psychology literature on self-justification. Thus, a farmer faced with adverse market or personal signals may obstinately refuse to quit rather than admit that his efforts and sacrifices (or those of his parents) have been in vain. A manager may keep throwing good money after bad on a doomed project (as in the original experiments of Staw (1976)). Others will keep accumulating wealth, professional achievements, political or religious activism, not so much for the marginal product of the later investments but to preserve the value of earlier ones –that is, to safeguard or strengthen the belief (true or false) that these assets will bring happiness over the course of their lifetime, or a favorable fate in some hereafter.

The escalating commitment result relies on $V_{23} > 0$, meaning that people have a higher demand for optimistic beliefs when they have more at stake. This assumption has substantial empirical support. Pyszczinsky (1982) found that lottery participants rated the prize as more desirable, the greater their perceived chance of winning it; Kay et al. (2002) found similar

outcomes among political partisans for electoral outcomes and among students for changes in tuition. Kunda (1987) had subjects read a (bogus) medical article linking cumulated caffeine consumption to risks of fibrocystic disease and breast cancer. Among the women, heavy coffee drinkers judged the information to be significantly less credible than light drinkers, whereas the men (who were not "at risk" and thus served as a control group) showed no such difference. Best known is the "Stockholm syndrome", in which hostages come to see their captors in a favorable light, most plausibly so as to maintain hope that they will not harm them.

4) Uncertain values. The overall (ex ante) probability of investment $\bar{x}$ is hill-shaped with respect to ρ : intuitively, investing in self-reputation has a low payoff when the prior is low, and is not needed when it is already high (provided $\bar{\rho} < 1$). This means, first, that identity-affirming behaviors are characteristic of people with unsettled preferences and values: hence the zeal of the new convert (religious or political), the nationalism of the recent immigrant (towards his new country or the old one) and the oppositional rituals of adolescents. People who are confident of "who they are", on the contrary, have no use for purely identity-affirming behaviors (they invest only if r_0 is large enough to justify the cost).[20]

Second, the model's predictions with respect to ρ can help understand and reconcile several disparate or even contradictory experimental findings concerning people's responses to manipulations of their self-image.

a) Substantial *identity threats* trigger large opposing responses aimed at restoring the damaged self-image –as occurs in the model when ρ is caused to fall below $\bar{\rho}$. In Maas et al. (2003), males subjects who were told by the experimenters that their score on a personality test was so atypical as to place them squarely in the female part of the distribution were subsequently much more likely than the control group to harass a female (but not a male) chat-line user by sending her pornographic images. This effect was further accentuated when she (a confederate) had previously described herself as a professionally ambitious feminist rather than a meek, family oriented traditionalist; it was also more pronounced, the more the subjects had initially self-rated themselves as masculine. Turning now from gender identity to "good person" identity, the same comparative-statics prediction can account for the *"transgression-compliance"* effect (e.g., Carlsmith and Gross (1969)): subjects who are led to believe that they have harmed someone (e.g., by administering painful electric shocks, or by carelessly ruining some of her work) show an increased willingness to later on accept requests to perform a good action, even though the requester is not their "victim" and does not even know about their "misdeed".

b) Moderate manipulations of an identity that is desirable but relatively fragile, on the other hand, are likely to lead to confirmatory rather than fighting responses –as occurs in the model

[20] In line with this "uncertainty principle," Adams et al. (1996) found that male subjects with strongly declared homophobia were in fact those who showed the most arousal in response to male homoerotic videos (with no difference from others subjects for heterosexual or female homoerotic materials).

when ρ changes marginally, starting from a value below $\tilde{\rho}$. Such is the case with the *"foot in the door"* effect (e.g., DeJong (1979)), in which freely accepting an initial request for a small favor raises the probability of accepting a more costly one in the future.[21] The model can similarly account for the debilitating impact of *"stereotype threat"* on test performance (Steele and Aronson (1995)). A social stereotype of female or African-American students as having a lower distribution of (say) comparative mathematical abilities than their male or White and Asian counterparts means precisely that society places a lower prior on their being a high type (with v now representing ability rather than taste, or a combination of both). Making gender or race subtly more salient before a test reminds the subjects of this statistical perception and thus (consciously or unconsciously) lowers their self-confidence. The equilibrium response to this decrease in ρ is (on average) to discourage academic-identity investment –in this case, effort and motivation to perform on the test.

C Identity and welfare: treadmill effect or empowerment?

While equilibrium behavior and most comparative results are quite general, relying only on Assumptions 1-3, the welfare implications of belief management depend critically on whether it reflects a demand for "consumable thoughts" or instrumental concerns.

1) Anticipatory utility and the treadmill effect.

Equations (4)-(6) lead to

$$\begin{aligned} W &= \rho x_H \left[(\delta_1 s_1 + \delta_2) v_H r_0 - c_0^H \right] + (1-\rho) x_L \left[(\delta_1 s_1 + \delta_2) v_L r_0 - c_0^L \right] \\ &\quad + [s_0 + \delta_1 s_1 + \delta_2] \overline{v} A_0. \end{aligned} \tag{17}$$

The last term is constant: although agents actively manage their self-views, this is a zero-sum game, by the law of iterated expectations.[22] As to the first two terms, they always (weakly) decrease as identity investments rise in response to a greater malleability of beliefs $1-\lambda$. This is immediate to see for an immutable characteristic like gender, race, or nationality: with $r_0 = 0$, there remains only a loss of $-\rho x_H c_0^H - (1-\rho) x_L c_0^H$. The result (a form of wasteful signaling) applies equally with an accumulable asset, however.

Most strikingly, an increase in his capital stock can also make the individual worse off.

[21] Conversely, an initial costly request, which most people turn down, decreases the probability of accepting a smaller one later on. In neither case are the results due to self-selection, since the probabilities being compared are the average compliance rates between the members of an experimental group (who get two requests) and those of a control group (who get only the second request).

[22] For welfare gains to arise, it must thus be that either: (i) agents' updating is at least partially naïve: when $a_0 = 1$, they do not properly correct for pooling by the v_L type, resulting in a departure from the martingale property of Bayesian beliefs. This additional form of malleability could easily be incorporated into the model (e.g., Bénabou and Tirole (2002)); or (ii) the consumption value of beliefs is nonlinear (and thus not purely anticipatory in the standard sense), as in some of the cases studied by Caplin and Leahy (2001) and Köszegi (2004).

Indeed, the condition for a no-investment equilibrium ($x_H = x_L = 0$),

$$\mathbf{V}(v_H, v_H, A_0 + r_0) - \mathbf{V}(v_H, \overline{v}, A_0) = (\delta_1 s_1 + \delta_2) v_H r_0 + (1-\lambda)\,\delta_1 s_1 (v_H - \bar{v})\, A_0 \leq c_0^H,$$

ceases to hold as A_0 crosses some threshold level. At that point investment jumps up discretely, resulting in a net welfare loss, by the same reasoning as above. More generally, the model yields a type of *treadmill effect:* higher levels of wealth, social status, professional achievements, etc., do no generate much of an increase in life satisfaction, or may even reduce it –and this precisely due to a self-defeating pursuit of the belief that these assets will bring happiness.[23]

Proposition 3 *In the anticipatory utility or self-image case,*
(1) An increase in the malleability of beliefs ($1-\lambda$) always reduces welfare.
(2) An increase in (per se valuable) capital A_0 can also make the individual worse off.
(3) The same holds for an increase in salience s_1.

A significant share of *advertising* involves playing up people's desires to achieve or affirm certain identities, by making more salient the benefits of being a high rather than a low type (raising s_1 with respect to beauty, wealth, social status, etc.) and targeting demographic subgroups with an insecure self-image, such as adolescents (ρ in the middle range). Our result shows that this can be quite effective in inducing consumers to purchase ($a_0 = 1$) and yet substantially lower their average welfare –and even social welfare, given that advertising is costly.

2) Willpower and the commitment value of identity

In the self-control version of the model, A_0 has no behavioral impact (unless some complementarity with a_0 is assumed), as seen from (9). The malleability of beliefs, on the other hand, now affects behavior both at $t=0$ and at $t=1$. Suppose for instance that when $\lambda = 1$ neither type invests at $t=0$ ($c_0^H > \delta_2 v_H r_0$), whereas for $\lambda < 1$ the equilibrium involves mixing ($0 < x_L < x_H = 1$).[24] The difference in intertemporal welfare, $W = E[\beta_0 U + V]$, is then

$$\Delta W = (1-\rho) x_L \left(\delta_2 v_L r_0 - \beta_0 c_0^L\right) + \rho\left(\delta_2 v_H r_0 - \beta_0 c_0^H\right) + (1-\lambda)\, E[\Delta V], \tag{18}$$

where the last term reflects the effects of belief management on date-1 behavior:

$$\begin{aligned} E[\Delta V] \;=\; & (1-\rho) x_L \left[F\left(\frac{\delta_1 c_1}{\delta_2 r_1 v_L}\right) - F\left(\frac{\delta_1 c_1}{\delta_2 r_1 \hat{v}(1)}\right)\right](\delta_2 v_L r_1 - \delta_1 c_1) \\ & -\rho\left[F\left(\frac{\delta_1 c_1}{\delta_2 \hat{v}(1) r_1}\right) - F\left(\frac{\delta_1 c_1}{\delta_2 v_H r_1}\right)\right](\delta_2 v_H r_1 - \delta_1 c_1). \end{aligned} \tag{19}$$

[23] Our is thus a different mechanism for treadmill effects from the traditional one, which is based on preferences or "aspirations" adapting to changes in consumption levels.

[24] This is without loss of generality: a similar reasoning applies for complete pooling (whether on 0 or on 1), with $\hat{v}(1)$ simply replaced by $\overline{v}$. Of course, the nature of the equilibrium, including the value of $\hat{v}(1)$, is endogenous and depends on the distribution $F(\beta_1)$. The proof of Proposition 4 takes this fixed-point aspect into account.

When $\lambda < 1$, the pooling which occurs at $t = 0$ *boosts* the v_L type's self-confidence and subsequent propensity to invest, but simultaneously *weakens* those of the v_H type. Since $a_1 = 1$ is always optimal (by (7)), the first effect leads to a welfare gain, the second to a loss. Thus, when $F(\cdot)$ is such that the support of $(\delta_1 c_1/\delta_2 r_1)/\beta_1$ is mostly concentrated in the interval $[v_L, \hat{v}(1)]$, meaning that the difficulty of the task and magnitude of the self-control problem are relatively moderate, there is a net gain from malleability. When they are more severe, so that the support is mostly concentrated in $[\hat{v}(1), v_H]$, there is a net a loss.[25]

Turning now to the direct contribution of date-0 behavior to ΔW, if β_0 is low enough that (say) the first two terms in (18) are positive, ex-ante efficient investments fail to occur in period 0 when $\lambda = 1$. The ability to manage one's beliefs ($\lambda < 1$) provides additional motivation for such investments, which then directly raise ΔW. When β_0 is near 1 such investments are a net cost, which only pays off in terms of improved decisions at $t = 1$ if $E[\Delta V]$ sufficiently positive.

Proposition 4 *In the self control case, more malleable beliefs (a lower λ) can raise welfare, by improving choices at $t = 1$ (when $E[\Delta V] > 0$) and/or at $t = 0$ (when $\Delta W > (1 - \lambda) E[\Delta V]$).*

Having completed the general positive and normative analysis of the model, we now turn to four economic applications.

III Taboos

While economists tend (at least, in their professional "identities") to view all activities as fungible or secular, that is, subject to trade-offs, most societies, religions and cultures hold, or at least declare, certain goods to be "priceless" or "sacred": life, justice, liberty, honor, love, friendship, one's children, democratic citizenship, religious faith, etc. (see, e.g., Durkheim (1925), Fiske and Tetlock (1997)).

It is thus considered highly immoral to attribute a monetary value to marriage, friendship or loyalty to a cause. Sexuality, death, body organs and military duty are not to be "commodified", nor are childbearing permits an acceptable policy for population control. Admittedly, such rules are often observed in the breach, and the boundaries between the secular and the sacred are evolving ones, as demonstrated by the changing attitudes toward life insurance (Zelizer (1999)), pollution permits, or, in certain places, legalized prostitution. Nonetheless, taboos often do bind, removing a number of activities from the traditional economic sphere or confining them to black markets. They also testify to widespread views that the mere existence of certain markets would be "contrary to human dignity" and harmful even to people who do not transact in them,

[25] When the two scenarios have equal probability, the net welfare effect is negative, since investment is more valuable when the true v is high. Thus, if $1/\beta_1$ is uniformly distributed on any subinterval of $[1, +\infty)$, the two bracketed terms in (19) are respectively proportional to $(1 - \rho) x_L [\hat{v}(1) - v_L]$ and $\rho [v_H - \hat{v}(1)]$, and thus equal.

because they would allow or "invite" comparisons and that, to use Fiske and Tetlock's (1997) memorable phrase, "to compare is to destroy". Yet what exactly is being destroyed by placing a monetary value on certain goods or activities, and how this damage occurs,

Taboos and sacred values are closely related to the preservation of identity, in the sense of protecting certain beliefs (or illusions), deemed vital for the individual or for society, concerning things one "would never do" and the "incommensurable" value of certain goods. To see this, let $v \in \{v_H, v_L\}$ represent the expected long-term value of an important state of being or social asset: freedom, bodily integrity, non-addiction, relationship to a person (child, spouse, friend) or to a more abstract entity (country, religion), with associated capital A_0. For the usual *anticipatory-utility* (including prospects for an afterlife) or *self-control* motives, people may want to be optimistic about v, resulting in a value function $V(v, \hat{v}, A_1)$ of the type studied earlier.

Suppose now that, at $t = 0$, an agent can find out the "sellout" price p at which he could exchange one unit of A_0 against money or other goods of known consumption value. *Ex ante,* the price could be high or low,

$$p = \begin{cases} p_H & \text{with probability} \quad z \\ p_L & \text{with probability} \quad 1-z \end{cases} . \tag{20}$$

The actual value may be learned, depending on the context, by checking what is being offered on a formal or informal market (for switching loyalties, selling one's vote, organ or children; for prostitution, fraud, crime, etc.) or by simply engaging in deliberate, "coldhearted" calculations about the costs and benefits of different courses of action.

To simplify the problem, let p_H be high enough and p_L low enough such that, if the agent does ascertain the price ($a_0 = 0$), he will always transact when $p = p_H$, reducing A_0 by one unit, and not transact when $p = p_L$.[26] In either case, he will later recall that he *entertained* the possibility of a transaction and *evaluated* whether maintaining his identity or dignity was "worth it" or not, and draw from this (with probability $1 - \lambda$) the appropriate inference about where his "true values" lie.

Investing in identity ($a_0 = 1$) thus consists here in upholding a rule never to not place a price on certain goods –staying away from markets where such transactions occur, not entertaining offers one may receive, and avoiding even "forbidden thoughts" of commensurability. The cost of doing so is the option value of the potential transactions thus foregone, so an individual of type $i = H, L$ will uphold the taboo if

[26] Formally, this is a dominant strategy for both types $i = H, L$, provided that $p_H > \mathbf{V}(v_H, v_H, A_0) - \mathbf{V}(v_H, v_L, A_0 - 1)$ and $p_L < \mathbf{V}(v_L, v_H, A_0) - \mathbf{V}(v_L, v_L, A_0 - 1)$. In the absence of such conditions, or with a more general price distribution, there may be two signals of an agent's type: whether he looked into the price and, if so, whether he transacted or not, given the price. We isolate here the first effect, which is the relevant one for the idea that certain things should remain "priceless".

$$\mathbf{V}(v,\hat{v}(1),A_0)-\mathbf{V}(v,\hat{v}(0),A_0-z)\geq zp_H, \tag{21}$$

with the same notation as usual.[27] This is clearly a special case of our model, with $r_0 = z$, $c_0 = zp_H$ and initial stock $A_0' \equiv A_0 - z$; therefore, all previous results apply directly. On the positive side, Propositions 1 and 2 show how taboos arise and are sustained, either universally (full-investment equilibrium) or predominantly by the more committed (mixing or separating equilibrium), how this depends on the initial strength of beliefs and how taboo-breaking by others can lead to reaffirmation or collapse.[28] On the normative side, Propositions 3 and 4 show how the welfare effect of taboos (absent externalities) depends critically on whether they reflect mental-consumption or self-control motives. In the first case, taboos generally lower ex ante welfare.[29] In the latter, they can increase it, but only under specific conditions involving priors and the severity of the self-control problem (or, under the intergenerational interpretation of the model, the misalignment of parental and child preferences).

IV Competing identities and dysfunctional behavior

We saw earlier how the single-asset model can be interpreted in "differential" form, as representing a tradeoff between two identity dimensions A and B whose relative value is uncertain and which are subject to resource rivalry at the investment stage. We analyze here a different kind of identity conflict, consumption rivalry, and show it can lead to highly dysfunctional behaviors.[30]

When time, geographical, legal or other exclusivity constraints (as with national or religious affiliations) create a potential tradeoff between reaping the future benefit from two identities, investing in one (say, B) damages the other (A), as it suggests that the individual may not value it that much. If he has substantial capital vested in A but the ultimate value of this identity is less "secure" than that of B, he may then refrain from even highly desirable investments in B and end up worse off as a result. We demonstrate here this mechanism using anticipatory utility or self-image, then discuss the more general case. We also make simplifying assumptions

[27] We assume that transacting without first finding out the price is either infeasible, or else unprofitable (due to the average "auction" price $zp_H+(1-z)p_L$ being too low). In writing the second term in (21) we took advantage of the linearity of $\mathbf{V}$ in A_1 under both the AU and the SC models (and their combination in Example 3). More generally, it would be $z\mathbf{V}(v,\hat{v}(0),A_0-1)+(1-z)\mathbf{V}(v,\hat{v}(0),A_0)$, which leaves all the results unchanged.

[28] See Section V for more details on peer effects. Because they involve the avoidance of normally valuable information, taboos are related to the strategic ignorance in Carrillo and Mariotti (2000) and Carrillo (2005), and especially to the rule-based behavior in Bénabou and Tirole (2004). There are, however, two important differences. On the demand side, imperfect willpower is here only one of several potential sources of motivated beliefs. On the supply side, it is the mere act of exploring the price to be gained from certain transactions, rather than the price thus revealed or whether the transaction is actually "consumed", that destroys the valued belief.

[29] Unless agents are sufficiently non-Bayesian, or the consumption value of beliefs is appropriately nonlinear: see footnote 22.

[30] The third type of interaction is correlation (of either sign) between v_A and v_B, which can lead to clusters of related behaviors, such as those indicative of the "disciplined self" and the "caring self" (Lamont (2000)).

under which A can be interpreted as the "traditional identity" and B as the "modern" one –for instance, in the context of farmers and workers faced with shifts brought about by globalization and technical change, or that of immigrants confronting the issue of assimilation.

(a) Modern identity. At $t = 0$, the agent decides whether to invest in B ($b_0 = 1$), at a cost c_B, type-independent for simplicity: acquiring new skills, mastering a new language and culture, socializing with an unfamiliar group, etc. The investment succeeds with probability $z \in (0, 1)$, in which case B_0 rises to $B_1 = B_0 + b_0 r_B$; it fails with probability $1 - z$ ($B_1 = B_0$), for instance because this is a new activity to which the agent may not be suited. The (per unit) value of B capital, on the other hand, is a known v_B. For instance, the monetary benefits of successfully integrating into the formal, majority-dominated labor market, of acquiring a degree or working in the more dynamic sectors of the economy are relatively easy to assess

(b) Traditional identity. There is no possibility of investment in A at $t = 0$. Thus A_0 corresponds either to a fixed trait (e.g., ethnicity) or to an asset that was accumulated in the past but can no longer be significantly augmented: long-held skills, connections to "the old country", etc.. Furthermore, the hedonic value of this stock is uncertain, since its benefits are of a more subjective and less quantifiable nature than, say, those of a wage premium: strength of personal values and commitments, long-run utility from family, morals, culture, religion, etc. Thus v_A equals v_H or v_L, with probabilities ρ and $1 - \rho$.

The timing is the same as before. At date 0, the agent receives the signal v_A, then chooses $b_0 \in \{0, 1\}$. At date 1, he recalls v_A with probability λ ($\hat{v}_A = v_A$), and otherwise looks to his past actions to form his sense of identity ($\hat{v}_A = \hat{v}(a_0)$). At date 2, he is aware of v_A (one could allow for uncertainty here as well) and, assuming full rivalry, chooses optimally between consuming either A or B, thus achieving $\max\{v_A A_2, v_B B_2\}$. To focus on the interesting case, suppose that: (i) *ex post,* the agent will consume B only if he had successfully invested in it,

$$v_B B_0 < v_L A_0 < v_H A_0 < v_B (B_0 + r_B), \tag{22}$$

so that A serves as a *"fallback"* option; (ii) *ex ante,* the expected return from investing in B is sufficiently high that, when beliefs are not malleable ($\lambda = 1$), such investment is optimal even for the agents who value A the most:

$$z (\delta_1 s_1 + \delta_2) [v_B (B_0 + r_B) - v_H A_0] > c_B. \tag{23}$$

When self-perception concerns are operative, however, both types will fail to make this efficient investment, as long as

$$z (\delta_1 s_1 + \delta_2) [v_B (B_0 + r_B) - v_L A_0] - (1 - z)\delta_1 s_1 (1 - \lambda) (\bar{v} - v_L) A_0 < c_B. \tag{24}$$

The first term is the standard economic return to investing, for an agent with relatively low valuation for A. The second term represents the *loss of identity* that is incurred (by either type) when doing so: with probability $1-\lambda$ such "betrayals" will signify to the individual that he does not care that much about A and therefore has only grim prospects to look forward to in case his investment in B does not work out.

On average, such savoring or affect-motivated identity management always ends up lowering welfare, as in the single-identity case. Indeed, while the nonlinear value function makes the model more complicated, one can exploit the basic intuition that not investing in B is effectively like investing in A to show that all the preceding results apply here as well.

Proposition 5 *Assume the AU specification, with (22)-(23). The individual invests (weakly) less in a known identity (B) when it will compete in the future with another one (A) of uncertain value. This is more likely to happen the higher* $A_0, 1-\lambda$ *and* s_1*, and it is always welfare reducing.*

These results directly relate to recent trends and controversies.

1) Resistance to structural change. Trade and technical change alter the relative payoffs to working in the modern, international sector and in traditional activities. The transition, which is risky and requires new skills and lifestyles, will be resisted if it is seen as de-valuing the old (rural, extended-family, blue-collar, etc.) identity.

2) Resistance to assimilation. Immigrants and their descendents experience strong tensions between integrating into Western societies and preserving their specific culture. This is particularly acute for the young, who are locally born and have citizenship but often do not feel British, German or French. Yet neither do they feel Pakistani, Turkish or Algerian, having little knowledge of the "old country" or its language. As seen earlier, it is in situations of uncertainty over one's own values that identity threats and investments become most relevant.[31] Laws and proposals such as the French ban on the veil or the Home Secretary's (2001) urging that minorities adopt British "norms of acceptability" and that newcomers take an oath of allegiance, study British history and culture and embrace "our laws, our values, our institutions" then elicit significant opposition from those whose who feel that complying would represent a betrayal of their own identity, culture or religion.[32]

In a related vein, it has been suggested that low educational achievement among African-Americans students may partly reflect a desire to maintain an "oppositional" ethnic identity. Austen-Smith and Fryer (2005) assess the evidence and model a form of "acting White" in which

[31] One can also relate to the results in Proposition 2 on the effects of A_0 and ρ the findings by Constant et al. (2006) that, among immigrants to Germany, the probability of assimilation decreases with age at arrival and with having had primary or secondary schooling in the country of origin.

[32] See Hoge (2002). Here again, self-perceived intentions matter: infiltrated members of an extremist organization feel much less conflict in submitting to such requirements, pledges, dress codes, etc., because they know that their doing it really signals commitment to, rather than abandonment of, their chosen "values".

some minority students forsake educational investment in order to signal their loyalty to peers and neighbors: demonstrably low labor market prospects means that they are unlikely to one day leave without "giving back to the community". Proposition 5 has both important parallels (in outcomes) and differences (in the mechanisms involved) with their result.[33]

2) Destructive identity, discrimination and communitarianism. "Not investing in B" in order to safeguard A can also mean destroying productive B capital. This simply corresponds in the model to the case where $c_B < 0$, so that the costly action is now one that reduces B or prevents it from growing ($b_0 = 0$). In the events that shook the suburbs of French cities in 2005, the young rioters attacked and destroyed a number of schools and nursery schools, pharmacies and many cars, mostly in their in *own* communities.

It is also interesting to note two factors that can "tip" the equilibrium from one in which people optimally invest in B to one in which they self-defeatingly destroy those assets (i.e., affecting (24) while leaving (22) and (23) unchanged). The first is a lower perceived chance of success in those investments (z) or associated payoff (r_B). Thus, if minority youth become more pessimistic about their chances of mobility through education, or perceive, rightly or wrongly, that even with diplomas the jobs to which most of them will be able to aspire will be low paying ones, they will switch to the destructive-identity scenario, even when z and r_B remain high enough that investing in B (education, integration) would *still* make them better off in the long run. A second potentially important factor is the salience s_1 of the "alternative" A identity and the benefits anticipated from it –as with advertising in the single-identity case. This is where ideological manipulation and religious indoctrination may come into play, as well as the amplification mechanism of media coverage.

Finally, while we have focused here on the anticipatory-utility or self-image case, which is somewhat simpler and seems more appropriate to the applications just discussed, similar insights clearly apply when the demand for identity stems from a commitment problem. If the individual expects sufficient temptation to underinvest in A relative to B at $t = 1$, he will not invest in B at $t = 0$ even if it has a high return, and may even destroy B capital. Such a strategy serves not as a physical commitment (investment costs and returns at $t = 1$ are independent of the stocks) but as a *cognitive* one, aimed at *defining oneself* as an A-person rather than a B-person. From Proposition 4 we know that welfare may go up in this case, but need not.

[33] In our case the (stochastic) returns to education are common knowledge and there is no incentive to deceive others. Instead, the individual wants to sincerely believe, and thus tries to convince himself, that his community is very valuable to him (instead of his being valuable to them). Since this mechanism does not involve community enforcement of membership "payments" through the expulsion of defectors, the relevant community or identity capital can be far away, uncoordinated, or even virtual (e.g., native country, culture, religious faith).

V Peer effects and norm transgressions

People's social environment shapes their identity through two channels. First, interactions with family, peers or coworkers directly affects the payoffs (cost c_t, productivity r_t or value v) of investing in different identity-relevant assets. Second, they affect the two cognitive mechanisms involved in belief management, by exposing the individual to informative signals and cues (supply side) and by altering his incentives to view himself, and be viewed by others, as a certain kind of person (demand side). Our focus here is on the cognitive channel.

1. Social signaling. In addition to his self-image $\hat{v}$, an individual often also cares about others' perceptions $\hat{v}'$ of his type, resulting in a continuation value of the form $V(v, \hat{v}, A_1, \hat{v}')$. This may reflect affective concerns for social esteem, posterity, etc., or strategic ones, for instance if others are more likely to make investments that benefit him when $\hat{v}'$ is high –because their own expected return is correlated with v, or is higher if the agent himself invests (or exerts self-control).[34] Since others form their beliefs by observing behavior, $\hat{v}' = \hat{v}(a_0)$ and the expected value function playing the role of (15) is now

$$\mathbf{V}(v, \hat{v}, A_1) \equiv \lambda V(v, v, A_1, \hat{v}) + (1 - \lambda) V(v, \hat{v}, A_1, \hat{v}). \tag{25}$$

Thus, as long as $(v, \hat{v}, A_1) \longmapsto V(v, \hat{v}, A_1, \hat{v})$ satisfies Assumption 3, adding a social signaling concern is akin to amplifying the self-signaling motive (from V_2 to $V_2 + V_4$), and the whole analysis, positive and normative, carries over.

2. Responding to transgressions. We now examine how observing others' behavior, rather than being observed by them, influences behavior. We consider here a very simple, two-agent version of the basic model, with sequential moves.[35] The two agents need not be symmetrical. At date 0, j moves first, choosing a_0^j; then, after observing j's action, i makes his own choice a_0^i. These are just investments in specific assets and do not directly enter into the other agent's payoff (so, at this stage, $a_0^i = 1$ is not to be interpreted as aggressiveness, ostracism, etc.). The only link between the two individuals is that they are "similar", that is, their values (v^i, v^j) are affiliated. Let ρ_0 be the prior on i's type and ρ^+ and ρ^- the corresponding posteriors after observing j invest $(a_0^j = 1)$ or not invest $(a_0^j = 0)$. Under a monotonic strategy for j, and with positive affiliation between v^i and v^j,

$$\rho^- < \rho_0 < \rho^+, \tag{26}$$

since j's investing (say) is "good news" about her v^j and therefore also about v^i. Furthermore,

[34] These correspond to interpersonal interpretations of (9), with $\hat{v}$ replaced by $\hat{v}'$. In Rotemberg (1994), similar complementarities lead agents to "invest in altruism" (even unilaterally), thus altering their own preferences rather than their beliefs and external image, as is the case here.

[35] The case of simultaneous moves is more complicated, as it involves mutual informational spillovers between agents' actions, plus coordination of their expectations. Battaglini, Bénabou and Tirole (2005) provide a detailed analysis of peer interactions among agents with a self-control problem related to that in Example 2.

if one fixes x_L^j and x_H^j, which is legitimate if j does not observe i's action, or else has no doubt about his own values, then as the correlation of types increases, ρ^- decreases and ρ^+ increases (weakly). To derive how i's behavior is influenced by j's, we can then directly apply Propositions 1 and 2, with the initial belief ρ set to ρ^- or ρ^+. Proposition 2.2 also readily yields the comparative statics of average investment with respect to the degree of correlation, which just acts like a *mean-preserving spread* in ρ.[36]

These results can help understand the nature of identity threats coming from other persons or groups and how people deal with transgressors. "Deviant" behavior by peers (non investment) sends a negative signal about the value of the existing capital stock (anticipatory utility version) or that of motivation-sensitive future investments (imperfect willpower version). For example, members of an ethnic, religious or national community who mingle with "outsiders", or are not fully supportive of the group's positions, undermine others' sense of commitment to the common value. Or, as discussed by Akerlof and Kranton (2000), a woman in a construction job or a man wearing a dress threaten masculine identity –more specifically, in our model, men's beliefs about abilities "only they" have, or attractions they "could never" have. When such transgressions represent sufficiently bad news to an initially strongly held identity ($\rho^- < \overline{\rho} < \rho_0$), they will elicit a strong investment response, designed to *"repair"* the damaged self-view. When the initial identity was relatively weak, on the other hand ($\rho^- < \min\{\tilde{\rho}, \rho_0\} < \overline{\rho}$), transgressions will further *"sap morale"* and depress investment.

Focussing on the first case, many strong reactions to deviant behaviors can be partly understood as cognitive strategies.[37] First, the exclusion of mavericks from the group suppresses the undesirable *reminders* created by their presence: "out of sight, out of mind". That is, exclusion lowers λ. Second, ostracism or harassment can be a form of belief damage control by self-signaling: one must forego beneficial interactions with the excluded and expend resources or take risks to support norm enforcement (including punishing others who fail to enforce it). If those most truly committed to the group identity (v_H types) face lower costs in such activities, the sorting condition will hold for this less benign interpretation of a_0^i as well.

The bad news conveyed by a transgression is more threatening, the more similar the violator, that is, the more correlated the values *a priori.* The harshest moral condemnations and punishments are thus reserved for "insiders" who, by their words or acts, threaten a group's valued beliefs. The canonical example (so to speak) is apostasy. The Catholic Church long imposed excommunication on apostates, and tortured and executed heretics; the Sharia still prescribes that apostates should be put to death, lose their children and their property.

[36] As $\bar{x}(\rho) \equiv \rho x_H(\rho) + (1-\rho)x_L(\rho)$ is concave up to $\bar{\rho}$ then falls to 0, a (small) mean-preserving spread reduces it when starting from a prior $\rho_0 < \bar{\rho}$ and increases it when starting from $\rho_0 > \bar{\rho}$, provided ρ^- falls below $\bar{\rho}$.

[37] In addition to instinctive anger and contempt, hardwired or learned early on in life, which serve broad functional purposes as well. Such emotions also provide a suitably noisy "rationalization" for having excluded previous member, other than just censoring their messages.

VI Dignity and scapegoating in bargaining and group conflict

> If you cut the pay of all but the superperformers, you have a big morale problem. Everyone thinks they are a superperformer. (Head of human resources of a manufacturing company, in Truman Bewley, *Why Wages Don't Fall During a Recession)*

We consider here another set of economic and political applications of the model: how pride, dignity or wishful thinking about one's options ("keeping hope") lead individuals or groups to walk away from "reasonable" offers, try to shift blame for failure onto others, or take refuge in political utopias –leading to impasses and conflicts. Examples include trials, divorces, strikes, the scapegoating of minorities and certain wars. The importance of belief distortion in those phenomena is attested by field observers (e.g., Bewley (1999) in the context of labor relations, Woods et al. (2006) in that of war), as well as by recent experiments. In particular, Thompson and Loewenstein (1992) and Babcock et al. (1995) demonstrate how subjects in bargaining situations with common knowledge spontaneously generate, through self-serving processing and *recall* of the evidence, divergent beliefs about the fairness of their cause and wishful predictions of outcomes that, in turn, result in costly delays and failures to agree.

To capture these phenomena we consider a "partnership" between two individuals or groups –husband and wife, labor and management, majority and minority populations, etc. Each may be of high or low type, $i = H, L$, corresponding to ability, motivation, honesty, outside opportunities, etc. At date 0, the joint output or productivity of the partnership is revealed: it is either good or bad, $y \in \{y_B, y_G\}$, with $y_G > y_B$. The technology exhibits complementarity, in that $y = y_G$ if and only if $i = j = H$. The interesting case will then be when $y = y_L$, since this means that at least one of the parties is "to blame" for the low output –disappointing marriage, firm or economy, lost war, etc.

At the end of period 0, the two partners must decide whether to: (i) remain together, in which case they will continue to produce the same (expected) output in period 2 (the long run) and must bargain over how it will be shared; or (ii) split, in which case each type i will get a reservation value v_i, with $v_H > v_L$: producing in autarky, searching for a new match, or triggering a costly fight with the other side over the control of resources. In all that follows, we abstract from discounting ($\delta_1 = \delta_2 = 1$).

Let parameters be such that staying together is efficient for all teams, both balanced (HH or LL) and unbalanced (HL), but in the latter case a compensating transfer (or share of y_B exceeding 1/2) is needed to induce the more productive partner to stay:

$$y_G > 2v_H > \ y_B \ > v_H + v_L > 2v_L. \tag{27}$$

When bargaining and making their stay or quit decisions at the end of period 0, the two parties will be assumed to know not only the joint output y, but also each one's type. Such

symmetric-information will make inefficient-breakdown results all the more interesting, and allow us to provide a formal model of the Babcock et al. (1995) findings described above. In keeping with the rest of our self-inference based theory, we further assume that, at date 1 :

(i) Whereas the level of joint output y is "hard" data that is easy to remember and verify, individuals's separate contributions to it –their types v– represent soft, unverifiable information, which later on is only imperfectly recalled.[38] Indeed, it would always be more pleasant, *ceteris paribus,* to "recall" that one was the competent and honest partner and the other was entirely to blame for the team's poor performance. For notational simplicity we shall take here the recall probability of the v's to be $\lambda = 0$, but this is inessential.

(ii) Individuals experience anticipatory feelings from their long-run (date-2) income or consumption prospects, with savoring coefficient s_1, common to both for simplicity. Alternatively, they may derive utility from their self-view about their talent or usefulness to society; this slight variant leads to similar results.

We formalize the bargaining process over future output as a standard Nash demand game. At $t = 0$, with full and symmetric information, players 1 and 2 simultaneously make demands for shares θ_1 and θ_2 of future output, y.[39] If $\theta_1 + \theta_2 \leq 1$ each gets what they asked for, whereas if $\theta_1 + \theta_2 > 1$ the negotiation breaks down and the pair dissolves. We assume that offers are later remembered (having been formally recorded, submitted to an arbitrator, etc.), but the key results are similar when they are not.

We look for a symmetric, pure-strategy equilibrium, with shares $\theta_H^* > 1/2 > \theta_L^*$ for the high and low valuation types respectively in an unbalanced partnership, and a common share $1/2$ in a balanced one. Finally, we restrict out-of equilibrium beliefs as follows. Let $\Theta \equiv \{\theta_L^*, 1/2, \theta_H^*\}$ the set of equilibrium offers. For $\theta_i \in \Theta$ and $\theta_j \notin \Theta$, player i is presumed to have played on the equilibrium path, which is sufficient to tie down both players' types. If θ_i and θ_j are both in Θ but are jointly inconsistent with equilibrium, on the other hand, then: (i) if $\theta_i = \theta_j$ both players are considered equally likely to have deviated, resulting in $\hat{v}_i = \hat{v}_j = \bar{v} \equiv (v_H + v_L)/2$; (i) if $\theta_i > \theta_j$, then $\hat{v}_i = v_H$ and $\hat{v}_j = v_L$; this is in the spirit of standard equilibrium refinements, since it is always the strong type who has less to lose from breaking up the match.

Let us first observe that in any equilibrium with agreement, it must be that the shares demanded by both sides sum to 1; otherwise, either party can ask for ε percent more and gain

[38] Given the same information, subjects in bargaining situations systematically recall more of the evidence that favors their own side, event when roles are exogenously determined (Thompson and Loewenstein (1992)). In dictator games, they take advantage of contextual ambiguity to "persuade" themselves that they deserve more than what they judge to be the fair share when making allocations between other people (Konow (2000)).

[39] A larger share may correspond to a monetary transfer, a control right (e.g., regional autonomy, child custody), a prestigious position or a new performance measurement system that will favorably alter the sensitivity of income shares to individual contributions. We treat the allocation of period-0 output as sunk (e.g., shared *ex ante* on a 50-50 basis, before types are revealed). Since expected output is equal in both periods, allowing initial resources to be part of the bargaining would simply amount to doubling the size of the pie.

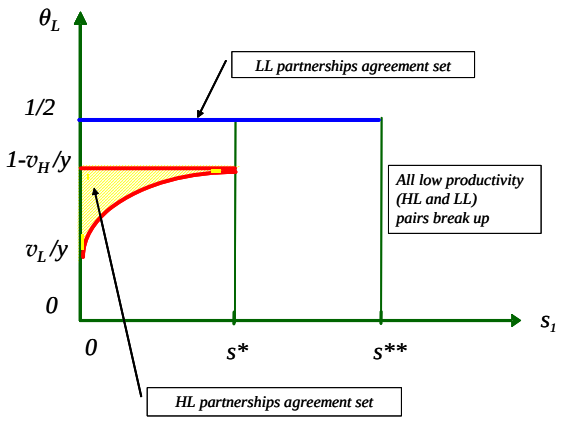

Figure III: Bargaining Sets and Breakdown Regions (for $s^* < s^{**}$)

$(1+s_1)\varepsilon y$, since the team will still stay together. For the same reason, downward deviations by either type (asking for less than the equilibrium share) are never profitable. The binding constraints will thus correspond to upward deviations.

Since $(1+s_1)y_G/2 > (1+s_1)v_H > (v_H + s_1\hat{v})$ for any feasible value of $\hat{v}$, matched strong partners (HH) always stay together, sharing output equally. The interesting case is that of low-productivity pairs, $y = y_B$. Consider first bargaining in an unbalanced (HL) team. For the H type to be satisfied with his share, it must be that:

$$\theta_H^* y_B \geq v_H. \tag{28}$$

Otherwise he could ask for more, which would break up the team while maintaining his posterior belief $\hat{v} = v_H$ and achieving $(1+s_1)v_H > (1+s_1)\theta_H^* y_B$. Next, for the weak partner (L type) to accept the bargain, it must be that:

$$(1+s_1)\,\theta_L^* y_B \geq v_L + s_1\bar{v}, \tag{29}$$

otherwise he could deviate and demand θ_H^* (mimicking the strong partner), thus achieving (and savoring at $t=1$) the posterior self-view $\hat{v} = \bar{v}$, even though his true outside option is only v_L. Other deviations to $\theta' > \theta_L$ with $\theta' \neq \theta_H$ would still identify him as the weak type, $\hat{v} = v_L$, and be *a fortiori* unprofitable under (29).

The set of mutually agreeable sharing rules $(\theta_L^*, 1-\theta_L^*)$ is thus defined by

$$\frac{v_L + s_1\bar{v}}{1+s_1} \;\leq\; \theta_L^* y_B \;\leq\; y_B - v_H. \tag{30}$$

As illustrated in Figure III, it shrinks as identity concerns increase, up to

$$s^* \equiv \frac{y_B - v_L - v_H}{v_H + \bar{v} - y_B} \tag{31}$$

when $y_B < v_H + \bar{v}$ (otherwise, let $s^* \equiv +\infty$). Beyond this critical threshold a *bargaining impasse arises,* in spite of gains from trade and symmetric information. Intuitively, a higher s_1 makes the loss of self-image involved in "admitting blame" more costly for the L type, who then requires a higher share θ_L^* to be compensated. At some point this becomes more than the H type is willing to grant given his outside option, and no agreement can be reached. The two parties then split (or fight), with each side getting $v_i + s_1\bar{v}$. Thus, once again, there is *in fine* no net gain in self-esteem or anticipatory utility, only a destruction of surplus.

We next turn to bargaining in an LL team. By asking for a share $\theta' > 1/2$, either side can break up the match and achieve self image v_H (by either of our refinements). Therefore, the partnership remains sustainable only if $(1 + s_1)\, y_B/2 \geq v_L + s_1\bar{v}$, or $s_1 \leq s^{**}$, where

$$s^{**} \equiv \frac{y_B - 2v_L}{2v_H - y_B}. \tag{32}$$

Otherwise the match is dissolved, as each side seeks to convince himself that he is better than the other, even though in reality both are equally bad; see again Figure III.

Finally, we can obtain a further result by naturally linking joint output to individual productivities. Consistent with our earlier assumptions, let HL and LL pairs both produce $y_B = \Phi v_L$, where Φ is such that (27) holds. It is then simple to verify that, as v_H/v_L rises, s^* and s^{**} both decrease and (30) becomes more stringent.

Proposition 6 *(1) For $s_1 \leq s^*$, unbalanced (HL) low-output partnerships successfully negotiate, splitting resources according to any sharing rule θ_L^* satisfying (30); this agreement range shrinks with s_1. For $s_1 > s^*$, the match is inefficiently destroyed*
(2) For $s_1 \leq s^{**}$, balanced (LL) low-output partnerships LL successfully negotiate, splitting resources equally. For $s_1 > s^{**}$, the match is inefficiently destroyed
(3) Let $y_B = \Phi v_L$. For any s_1, the bargaining set shrinks and both types of impasses become more likely, the greater the inequality v_H/v_L between high and low types' productivities.

Our model of bargaining with malleable beliefs thus identifies a new and potentially important limit to the achievement of Coasian deals, namely the preservation of dignity, pride, or "hope" about the future. It also leads to testable predictions, as both salience s_1 and the productivity differential v_H/v_L can be manipulated experimentally, and the latter at least could even be measured empirically in actual bargaining contexts.

VII Conclusion

We developed in this paper a simple but flexible framework for analyzing a broad class of beliefs which people value and invest in, with important economic implications. The model also offers a unified account of many disparate phenomena documented by psychologists and experimental economists; others, such as endowment effects, could be fairly easily obtained. Rather than restate here the paper's results, we will single out the two that are most novel and, having been explored only in their simplest form, suggest avenues for further research. The first is that of *sacred values and taboos,* where our framework offers a way of bringing the debate over markets and morals into the realm of formal analysis. The second concerns the role, in bargaining and other distributive conflicts, of *endogenously arising* self-serving beliefs linked to pride, dignity or wishful thinking. Potential applications include the design of contracts or organizations and the political economy of reforms.

Appendix: Proofs

Proof of Proposition 1. The difference between the two types' incentives to invest in (16) is

$$\Delta \equiv \int_{v_L}^{v_H} \left[\int_{A_0}^{A_0+r_0} \mathbf{V}_{13}\left(x, \widehat{v}\left(1\right), z\right) dz + \int_{\widehat{v}(0)}^{\widehat{v}(1)} \mathbf{V}_{12}\left(x, y, A_0\right) dy \right] dx + c_0^L - c_0^H. \tag{A.1}$$

If $V_{12} = 0$ (as with anticipatory utility) then $\Delta > 0$, so *any* equilibrium must have $x_L\left(1 - x_H\right) = 0$. When $V_{12} > 0$ the same holds provided $\hat{v}(1) \geq \hat{v}(0)$, but since those beliefs are endogenous we must make monotonicity a requirement. The possible equilibrium configurations are then:

(a) No investment: $x_H = x_L = 0$, hence $\hat{v}(0) = \overline{v}$ and $\hat{v}(1) = v_H$, with

$$\mathbf{V}\left(v_H, \overline{v}, A_0\right) \geq \mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - c_0^H. \tag{A.2}$$

(b) Randomization by v_H : $1 > x_H > x_L = 0$, hence $\hat{v}(1) = v_H$ and $v_L < \hat{v}(0) < \overline{v}$, with

$$\mathbf{V}\left(v_H, \hat{v}\left(0\right), A_0\right) = \mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - c_0^H.$$

(c) Separation: $1 = x_H > x_L = 0$, hence $\hat{v}(1) = v_H$ and $\hat{v}(0) = v_L$, with

$$\mathbf{V}\left(v_H, v_L, A_0\right) \leq \mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - c_0^H, \tag{A.3}$$

$$\mathbf{V}\left(v_L, v_L, A_0\right) \geq \mathbf{V}\left(v_L, v_H, A_0 + r_0\right) - c_0^L. \tag{A.4}$$

(d) Mixing by v_L: $1 = x_H > x_L > 0$, hence $\hat{v}(0) = v_L$ and $\overline{v} < \hat{v}\left(1\right) < v_H$, with

$$\mathbf{V}\left(v_L, v_L, A_0\right) = \mathbf{V}\left(v_L, \hat{v}\left(1\right), A_0 + r_0\right) - c_0^L. \tag{A.5}$$

(e) Full investment $x_H = x_L = 1$, hence $\hat{v}(0) = v_L$ and $\hat{v}\left(1\right) = \overline{v}$, with

$$\mathbf{V}\left(v_L, v_L, A_0\right) \leq \mathbf{V}\left(v_L, \overline{v}, A_0 + r_0\right) - c_0^L. \tag{A.6}$$

We can first rule out equilibria of type (b), in which type v_H randomizes: since $\mathbf{V}_2 > 0$, the no-investment equilibrium also exists if an equilibrium of type (b) exists. Furthermore, since $V\left(v, \bar{v}, A_0\right) > V\left(v, \hat{v}\left(0\right), A_0\right)$ for all v, both types are better off in the no-investment equilibrium, so we can apply the Pareto criterion in order to select the policy equilibrium. For the same reason, we can rule out the separating equilibrium (type (c)) whenever it coexists with the no-investment equilibrium (type (a)).

We now show that there exists a unique equilibrium, which involves no investment when (A.2) holds and, when this condition fails, separation, randomization by v_L or full investment, depending respectively on whether (A.3)-(A.4), (A.5) or (A.6) holds.

1) If $\mathbf{V}(v_H, v_L, A_0) \geq \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H$, it is a dominant strategy for both types not to invest, so $x_H = x_L = 0$ for all ρ, or equivalently $\bar{\rho} \equiv 0$.

2) Assume now that $\mathbf{V}(v_H, v_L, A_0) < \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H$. Because $\bar{v} \simeq v_L$ for ρ small, the no-investment regime (a) cannot prevail for ρ small. More generally, it obtains whenever $\rho \geq \bar{\rho}$, where $\bar{\rho} > 0$ is defined by

$$\mathbf{V}(v_H, \bar{\rho} v_H + (1 - \bar{\rho}) v_L, A_0) \equiv \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H \tag{A.7}$$

if this equation has a solution in $(0, 1)$ and to 1 otherwise. For $\rho < \bar{\rho}$ we have $x_H = 1$ from the previous taxonomy and the Pareto-dominance assumption.

If (A.4) holds, the equilibrium is separating: $x_H = 1$ and $x_L = 0$. By contrast, if $\mathbf{V}(v_L, v_L, A_0) < \mathbf{V}(v_L, v_H, A_0 + r_0) - c_0^L$, the v_L type must invest with positive probability. If (A.6) holds there can be no solution to (A.5) with $x_L < 1$, so the only equilibrium is full investment on $[0, \bar{\rho})$. If (A.6) is reversed, on the other hand, it involves mixing: by (14),

$$\hat{v}(1) = \frac{\rho}{\rho + (1 - \rho) x_L} v_H + \frac{(1 - \rho) x_L}{\rho + (1 - \rho) x_L} v_L, \tag{A.8}$$

and by (A.5) this expression must be independent of ρ. Thus, $x_L = (\gamma - 1)/(1/\rho - 1)$, where $\gamma = 1/\hat{\rho}(1) > 1$ is also a constant. If $(\gamma - 1)/(1/\bar{\rho} - 1) < 1$, then the v_L type mixes over all of $[0, \bar{\rho}]$; if $(\gamma - 1)/(1/\bar{\rho} - 1) \geq 1$, define $\tilde{\rho}$ by $(\gamma - 1)\tilde{\rho}/(1 - \tilde{\rho}) \equiv 1$ or, equivalently,

$$\mathbf{V}(v_L, v_L, A_0) = \mathbf{V}(v_L, \tilde{\rho} v_H + (1 - \tilde{\rho}) v_L), A_0 + r_0) - c_0^L. \tag{A.9}$$

Then $x_L \in (0, 1)$ for $0 < \rho < \tilde{\rho}$ and $x_L = 1$ for $\rho \geq \tilde{\rho}$. ■

Proof of Proposition 2. (1)(i) When λ decreases, each type v's incentive to invest, $\mathbf{V}(v, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v, \hat{v}(0), A_0)$, increases: indeed, by (15), its derivative with respect to $1 - \lambda$ is

$$V(v, \hat{v}(1), A_0 + r_0) - V(v, v, A_0 + r_0) + V(v, v, A_0) - V(v, \hat{v}(0), A_0) \geq \int_{\hat{v}(0)}^{\hat{v}(1)} V_2(v, x, A_0) dx > 0,$$

where the first inequality follows from the assumption $V_{23} \geq 0$. Consequently, the no-investment region shrinks, $\bar{\rho}$ increases, $\tilde{\rho}$ rises and $\hat{v}(1)$ decreases in the mixing equilibrium: investment increases (weakly) for each type, at any value of ρ.

(ii) It is easily verified from (A.7), (A.8) and (A.9) that a decrease in c_0^H increases $\bar{\rho}$ while a decrease in c_0^L decreases $\tilde{\rho}$ and reduces $\hat{v}(1)$ in the mixing region, thus increasing x_L. Thus, again investment unambiguously increases.

(iii) and (iv). In the AU case,

$$\mathbf{V}(v, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v, \hat{v}(0), A_0) = \delta_1 s_1 \left[\lambda v r_0 + (1 - \lambda)[\hat{v}(1)(A_0 + r_0) - \hat{v}(0) A_0] + \delta_2 v r_0\right.$$

rises with $s_1$and A_0. The rest of the proof follows the steps of part (i).

(2) The result is obvious when $x_L(\tilde{\rho}) = 0$ (separating equilibrium), since $x_L(\rho) \equiv 0$ in that case. When $x_L(\tilde{\rho}) > 0$ (equilibrium with randomization), it follows from the fact that $\widehat{v}\,(1)$ and therefore $\hat{\rho}(1) = \rho/\left[\rho + (1-\rho)x_L(\rho)\right]$ must remain constant over $[0, \tilde{\rho}]$. ■

Proof of Propositions 3 Consider (17). If $(\delta_1 s_1 + \delta_2)\, v_L r_0 \geq c_0^L$, it is a dominant strategy for both types to invest, so $x_H = x_L = 1$ and changes in λ do not affect behavior, nor W. If $(\delta_1 s_1 + \delta_2)\, v_H r_0 < c_0^H$, then W decreases with both x_H and x_L, so a decrease in λ can only (weakly) lower welfare. Finally, when $(\delta_1 s_1 + \delta_2)\, v_H r_0 - c_0^H \geq 0 > (\delta_1 s_1 + \delta_2)\, v_L r_0 - c_0^L$, type v_H always invests ($x_H = 1$); hence λ can only affect x_L, and any increase in x_L reduces welfare. The proof for small changes in A_0 around the no-investment threshold (given by the equation following (17)) is similar, since the direct effect on the last term in (17) is infinitesimal, whereas the jump in x_H (and possibly x_L) is discrete. ■

Proof of Proposition 4 The proof is by construction of an appropriate mixed equilibrium. Let us choose $\beta^* \in (0,1)$ such that $\beta^* \delta_2 r_1 \bar{v} < \delta_1 c_1 < \beta^* \delta_2 r_1 v_H$. Next, define $v^* \in (\bar{v}, v_H)$as $v^* \equiv (1/\beta^*)\,(\delta_1 c_1/\delta_2 r_1)$ and $x_L \in (0,1)$ by

$$\hat{\rho}(1) \equiv \frac{\rho}{\rho + (1-\rho)(1-x_L)} = \frac{v^* - v_L}{v_H - v_L}. \tag{A.10}$$

Suppose now that $F(\beta)$ puts mass 1 on β^*; by continuity, the arguments below will continue to hold when the mass is close enough to 1. By (8), the agent invests at $t = 1$ when $\hat{v} \geq v^*$. As to (A.10), it means that if the v_L type mixes at $t = 0$ with probability x_L, the posterior following $a_0 = 1$ is exactly v^*, inducing $a_1 = 1$ for both types. Next, choose c_0^H and c_0^L such that mixing with probability x_L defined by (A.10) is indeed the equilibrium:

$$c_0^H < \delta_2 r_0 v_H + (1-\lambda)\,(\delta_2 r_1 v_H - \delta_1 c_1) = \mathbf{V}(v_H, v_H, A_0 + r_0) - \mathbf{V}(v_H, \bar{v}, A_0), \tag{A.11}$$

$$c_0^L \equiv \delta_2 r_0 v_L + (1-\lambda)\,(\delta_2 r_1 v_L - \delta_1 c_1) = \mathbf{V}(v_L, v^*, A_0 + r_0) - \mathbf{V}(v_L, v_L, A_0). \tag{A.12}$$

Compared to the equilibrium that prevails when $\lambda = 1$, in which $\hat{v} = v$ always, this yields a gain in $E[V]$ given by (19) but with the loss term equal to zero; hence a positive contribution to welfare.

Turning now to period 0, in order for the equilibrium with $\lambda = 1$ to be one where neither type invests in spite of the fact that choosing $a_0 = 1$ would be be ex ante efficient for both (making the first two terms in (18) positive), if suffices that

$$\beta_0 c_0^L < \delta_2 v_L r_0 < \delta_2 v_H r_0 < c_0^H. \tag{A.13}$$

Compatibility with (A.11)-(A.12) requires that

$$\begin{aligned}(1-\lambda)\left(\delta_2 r_1 v_H - \delta_1 c_1\right) &> c_0^H - \delta_2 r_0 v_H > 0,\\ (1-\lambda)\left(\delta_2 r_1 v_L - \delta_1 c_1\right) &< \left(1/\beta_0 - 1\right)\delta_2 v_L r_0,\end{aligned}$$

neither of which contradicts any other condition. ■

Proof of Proposition 5 Given (22) and (23), the intertemporal utility of an agent of type $v_A \in \{v_H, v_L\}$ who starts with stocks (A_0, B_0) and chooses $b_0 \in \{0,1\}$ is:

$$\begin{aligned}&\tilde{W}(v_A, A_0, B_0, b_0) \equiv b_0 z\left(\delta_2 + \delta_1 s_1\right) v_B\left(B_0 + r_B\right)\\ &+\left(1-b_0\right)\left[\delta_2 v_A + \delta_1 s_1\left(\lambda v + (1-\lambda)\hat{v}_A(1-b_0)\right)\right] A_0\\ &+b_0(1-z)[\delta_2 v_A + \delta_1 s_1(\lambda v + (1-\lambda)\hat{v}_A(b_0))]A_0 - b_0 c_B. \qquad \text{(A.14)}\end{aligned}$$

Let us now define the variables $a_0 \equiv 1 - b_0$, $R_0 \equiv zA_0$ and the functions:

$$\begin{aligned}U(v_a, a_0, A_0; B_0) &\equiv \left(1-a_0\right)\left[z\left(\delta_2 + \delta_1 s_1\right) v_B\left(B_0 + r_B\right) - c_B\right],\\ V(v, \hat{v}, A_1) &\equiv \left(\delta_2 v_A + \delta_1 s_1 \hat{v}_A\right) A_1,\\ \mathbf{V}(v, \hat{v}, A_1) &\equiv \lambda V(v, v, A_1) + (1-\lambda) V(v, \hat{v}, A_1),\end{aligned}$$

It is then easy to see that (A.14) can be rewritten as

$$W(v_A, A_0, B_0, a_0) = U(v_a, a_0, A_0) + \mathbf{V}(v_A, \hat{v}_A(a_0), A_0(1-z) + a_0 R_0) \qquad \text{(A.15)}$$

and that the function U satisfies $U_3 = 0$, hence Assumption 1, while V is exactly the same as in (4) and therefore satisfies Assumption 3. Thus, although $U(v_a, a_0, A_0; B_0)$ and $\mathbf{V}(v, \hat{v}, A_1)$ no longer individually correspond to the date-zero flow payoffs and date-1 expected value function (e.g., U includes payoffs received at dates 1 and 2), their sum still defines the agent's objective function, with the only change with respect to the one dimensional problem being a minor one in the "fictitious" law of motion for A_t, which is now $A_1 = A_0(1-z) + a_0 R_0$. The depreciation" term in $1-z$ will not change anything (qualitatively), while the fact that the return $R_0 = zA_0$ now increases with the initial stock will only reinforce the fact that investment increases with A_0. Thus, the agent will invest at $t = 0$ if and only if

$$\mathbf{V}(v_A, \hat{v}(1), A_0(1-z) + R_0) - \mathbf{V}(v_A, \hat{v}(0), A_0(1-z)) \geq c_0, \qquad \text{(A.16)}$$

where $c_0 \equiv c_B - z\left(\delta_2 + \delta_1 s_1\right) v_B\left(B_0 + r_B\right)$ is now the same for both types. All results in Proposition 1 and all those in Proposition 2 pertaining to the anticipatory utility case thus remain unchanged. In particular, equilibrium generally results in excessive "investment" in A, which mean suboptimally low investments in B. ■

REFERENCES

Adams, H. Wright, L., and B. Lohr (1996) "Is Homophobia Associated With Homosexual Arousal?" *Journal of Abnormal Psychology,* 105(3),440-445.

Akerlof, G., and W. Dickens, (1982) "The Economic Consequences of Cognitive Dissonance." *American Economic Review,* 72(3), 307–319.

Akerlof, G., and R. Kranton (2000) "Economics and Identity," *Quarterly Journal of Economics*, 115, 716-753.

—— (2002) "Identity and Schooling: Some Lessons for the Economics of Education," *Journal of Economic Literature*, 40(4), 1167-1201.

—— (2005) "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19, 9–32.

Austen-Smith, D., and R. Fryer (2005) "An Economic Analysis of "Acting White"," *Quarterly Journal of Economics*, 120(2): 551–583.

Babcock, L., Loewenstein, G., Issacharoff, S. and Camerer, C. (1995) "Biased Judgments of Fairness in Bargaining**,"** *American Economic Review,* 85(1), 1337-1343.

Basu, K. (2006) "Identity, Trust and Altruism: Sociological Clues to Economic Development," Cornell University mimeo, April.

Battaglini, M., Bénabou, R., and J. Tirole (2005) "Self-Control in Peer Groups," *Journal of Economic Theory*, 123: 105–134.

Baumeister, R. (1986) *Identity: Cultural Change and the Struggle for Self.* Oxford: Oxford University Press.

Becker, E. (1973) *The Denial of Death*, New York: Free Press.

Becker, G. and K. Murphy (2000) *Social Economics: Market Behavior in a Social Environment.* Bellknap Press.

Bem, D. J. (1972). "Self-Perception Theory," in L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 6, 1-62. New York: Academic Press.

Bénabou, R. and J. Tirole (2002) "Self Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117(3): 871–915.

—— (2004) "Willpower and Personal Rules," *Journal of Political Economy*, 112(4): 848–886.

—— (2006a) "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics*, 121(2), 699-746.

—— (2006b) "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652-1678.

Bernheim, D. (1994) "A Theory of Conformity," *Journal of Political Economy*, 102(5), 842–877.

Bernheim, D. and R. Thomadsen (2005) "Memory and Anticipation," *The Economic Journal*, 115, 271–304.

Bewley, T. (1999) *Why Wages Don't Fall During a Recession.* Harvard University Press.

Benjamin, D., Choi, J. and J. Strickland (2006) "Social Identity and Preferences," Dartmouth College mimeo, September.

Bisin, A. and T. Verdier (2000) ""Beyond The Melting Pot": Cultural Transmission, Marriage, And The Evolution Of Ethnic And Religious Traits," *Quarterly Journal of Economics,* 115(3), 955-988.

Bodner, R. and D. Prelec (2003) "Self-signaling and Diagnostic Utility in Everyday Decision Making," in I. Brocas and J. Carrillo eds. *The Psychology of Economic Decisions. Vol. 1: Rationality and Well-being*, Oxford University Press.

Brunnermeier, M. and J. Parker (2005) "Optimal Expectations," *American Economic Review*, 95, 1092–1118.

Caplin, A., and J. Leahy (2001) "Psychological Expected Utility Theory and Anticipatory Feelings," *Quarterly Journal of Economics*, 116, 55–80

Carlsmith, J., and A. Gross (1969) "Some Effects of Guilt on Compliance," *Journal of Personality and Social Psychology*, 11, 232–239

Carrillo, J. (2005) "To Be Consumed with Moderation," *European Economic Review*, 49, 99–111.

Carrillo, J., and T. Mariotti (2000) "Strategic Ignorance as a Self Disciplining Device," *Review of Economic Studies*, 67(3), 529–544.

Cho, I.-K. and Kreps, David. (1987) "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics*, 102(2), 179–221.

Dana, J., Kuang, J., and R. Weber (2003) "Exploiting Moral Wriggle Room: Behavior Inconsistent with a Preference for Fair Outcomes." Carnegie Mellon Behavioral Decision Research Working Paper No. 349, June .

De Jong, H.W. (1979) "An Examination of Self-Perception Mediation of the Foot-in-the-Door Effect," *Journal of Personality and Social Psychology*, 37, 2221–2239.

Davies, J. (2004) "Identity and Commitment"' Tinbergen Institute Discussion Paper 055/2, University of Amsterdam.

Dessi, R. (2005) "Collective Memory, Social Capital and Integration," Université de Toulouse mimeo.

Durkheim, E. (1976) *The Elementary Forms of the Religious Life.* 2nd edition. London: Allen and Unwin (original work: 1925).

Fang, H. and Loury, G. (2005) ""Dysfunctional Identities" Can Be Rational," *American Economic Review,* 95(2), 104-111.

Festinger, L. and J. Carlsmith (1959) "Cognitive Consequences of Forced Compliance." *Journal of Abnormal and Social Psychology,* 58, 203–210.

Fiske, A., and P. Tetlock (1997) "Taboo Trade-offs: Reaction to Transactions that Transgress the Spheres of Justice," *Political Psychology*, 18, 255–297.

Fryer, R. and M. Jackson (2003) "Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making," NBER Working Paper 9579, March.

Geanakoplos, J., Pearce, D. and E. Stacchetti (1989) "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60–79.

Gul, F. (1991) "A Theory of Disappointment Aversion" *Econometrica,* 59(3), 667-686.

Hill, C. (2006) "What The New Economics Of Identity Has To Say To Legal Scholarship," University of Minnesota Legal Studies Research Paper No. 05-46.

Hoge, W. (2002) "Britain's Nonwhites Feel Un-British, Report Says," *New York Times,* April 4.

Horst, U., Kirman, A. and M. Teschl (2006) "Changing Identity: The Emergence of Social Groups," University of British Columbia mimeo, May.

Kahneman, D., and P. (1997) "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics*, 112, 75–407.

Kay, A., Jimenez, M. and J. Jost (2002) "Sour Grapes, Sweet Lemons, and the Anticipatory Rationalization of the Status Quo," *Personality and Social Psychology Bulletin,* 9, 1300-1312.

Konow, J. (2000) "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review,* 90(4), 1072-1091.

Kopczuk, W., and J. Slemrod (2005) "Denial of Death and Economic Behavior," *Advances in Theoretical Economics*, 5(1). Article 5.

Köszegi, B. (2004) "Utility from Anticipation and Personal Equilibrium," U.C. Berkeley mimeo, June.

Kunda Z. (1987) "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories," *Journal of Personality and Social Psychology,* 53(4), 636-647.

Lamont, M. (2004) *The Dignity of Working Men: Morality and the Boundaries of Race, Class and Immigration.* Harvard University Press.

Landier, A. (2000) "Wishful Thinking and Belief Dynamics," MIT mimeo.

LeBoeuf, R. and Shafir, E. (2004) "Alternative Selves and Conflicting Choices: Identity Salience and Preference Consistency," Princeton University mimeo.

Loewenstein, G. (1987) "Anticipation and the Valuation of Delayed Consumption," *Economic Journal*, 97, 666–84.

Loewenstein, G. and Schkade D. (1999) "Wouldn't It Be Nice? Predicting Future Feelings" in D. Kahneman, E. Diener and N. Schwartz, eds. *Well-Being: Foundations of Hedonic Psychology.* New York, NY: Russel Sage Foundation.

Maas, A. Cadinu, M., Guarnieri, G. and Grasselli, A. (2003) "Sexual Harassment Under Social Identity Threats: The Computer Harassment Paradigm," *Journal of Personality and Social Psychology,* 85(5), 853-870.

Mazar, N., Amir, O. and D. Ariely (2006) "Mostly Honest: A Theory of Self-Concept Maintenance," MIT mimeo, December.

Oxoby, R. (2003) "Attitudes and Allocations: Status, Cognitive Dissonance and the Manipulation of Preferences," *Journal of Economic Behavior and Organization,*, 52(3), 365–385.

—— (2004) "Status, Cognitive Dissonance, and the Growth of the Underclass," *The Economic Journal,* 114(498), 727–749.

Piccione, M. and A. Rubinstein (1997) "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior,* 20, 3-24.

Pyszczynski, T. (1993) "Cognitive Strategies for Coping with Uncertain Outcomes," *Journal of Research in Psychology,* 16, 386-399.

Quattrone, G., and Tversky, A. (1984) "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion," *Journal of Personality and Social Psychology*, 46(2), 237-248.

Rabin, M. (1994) "Cognitive Dissonance and Social Change," *Journal of Economic Behavior and Organization*, 23, 177-194.

—— (1995) "Moral Preferences, Moral Constraints, and Self-Serving Biases," Berkeley Department of Economics Working Paper No. 95-241, August.

Rotemberg J. (1994) "Human Relations in the Workplace," *Journal of Political Economy,* 102, 684-718.

Shayo, M. (2005) "Nation, Class and Redistribution: Applying Social Identity Research to Political Economy," Princeton University mimeo, August.

Schelling, T. (1985) "The Mind as a Consuming Organ." In J. Elster (Ed.), *The Multiple Self.* New York: Cambridge University Press, 177-195.

Sen, A. (1985) "Goals, Commitment, and Identity," *Journal of Law, Economics and Organization,* 1(2), 341-355.

Smith, A. (1759) *The Theory of Moral Sentiments.* Amherst, NY: Prometheus Books.

Smith, J. (2005) "Reputation, Social Identity and Social Conflict," Princeton University mimeo, November.

Steele, C. and J. Aronson (1995) "Stereotype Vulnerability and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology,* 69, 797-811.

Thompson, L. and G. Loewenstein (1992) "Egocentric Interpretations of Fairness in Negotiation," *Organization Behavior and Human Decision Processes,* 51, 176-197.

Woods, K., Lacey, J. and W. Murray (2006) "Saddam's Delusions: The View from the Inside," *Foreign Affairs,* June.

Wichardt, P. (2005) "Why and How Identity Should Influence Utility," University of Bonn mimeo, November.

Young, P. (2006) "Self Knowledge and Self-Deception," John Hopkins University mimeo, November.

Zelizer V. (1997) Morals and Markets: The Development of Life Insurance in the United States. New York: Columbia University Press.